DOCUMENT RESUME

ED 336 405 TM 017 078

AUTHOR Luijten, Anton J. M., Ed.
TITLE Issues in Public Examinations: A Selection of the Proceedings of the 1990 IAEA Conference (16th, Maastricht, Netherlands, June 1990).
REPORT NO ISBN-90-5189-091-5
PUB DATE 91
NOTE 184p.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Change; Educational Policy; Elementary Secondary Education; Foreign Countries; *International Studies; National Programs; Objective Tests; Postsecondary Education; *Public Education; Testing Problems; Testing Programs; *Test Use
IDENTIFIERS China; Europe; Standard Setting; United States

ABSTRACT

This collection of 18 papers (selected from a total of 57 presented at a conference of the International Association for Educational Assessment) represents efforts by examining bodies and institutes to: improve the examination system and testing techniques; develop reliable instruments; and establish standards for public examinations. The papers are: "Public Examinations, an Introduction" (A. J. M. Luijten); (2) "Social and Educational Imperatives for Changing Examinations" (P. J. Black); (3) "The Estimation of Bias in Matriculation Examination Scores and a Proposal for Correction" (S. Cahan); (4) "Recording Public Examination Performance--Usefulness, Accuracy, Ownership" (J. S. Cook); (5) "Divergence of Interests: Problems of Implementing Reforms in an Established Public Examination System" (R. K. Crocker); (6) "Attributes and Applications of a Pro-Active Assessment System" (R. L. Ferguson); (7) "Social Justice, Public Interest and the SAT" (C. Fincher); (8) "Issues in Assessment of Non-Objective and Objective Examination Tasks" (J. Izard); (9) "Determining Equivalence among European Educational Qualifications" (S. Johnson); (10) "Examinations in Europe: Three Major Streams" (H. P. Kreeft); (11) "The Politics of Assessment" (D. L. Ongom); (12) "Why Assess? Evaluating the Assessment of the English National Curriculum" (C. G. Robinson); (13) "Subject Pair Comparison Studies in an International Context" (J. J. Thompson and others); (14) "The Use of Item Response Theory in the Dutch National Assessment Programme" (N. D. Verhelst and T. J. H. M. Eggen); (15) "The Rationale of a School-Based Centralized Examination System" (W. H. Wijnen); (16) "Measuring Educational Progress with a Longitudinal Testing System" (T. L. Willoughby and A. R. Bixby); (17) "The Wessex Project: A Modular Development in English Post-Compulsory Education" (J. Wilmut and R. G. Rainbow); and (18) "Experiments on General High School Completion Tests in China" (X. Yang). (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Issues in public examinations

A selection of the Proceedings of the 1990 IAEA Conference

Anton J.M. Luijten

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. VAN DEN HEUVEL RIJNDERS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM017078

BEST COPY AVAILABLE

Issues in public examinations

A selection of the Proceedings of the 1990 IAEA Conference

Anton J.M. Luijten (editor)

1991, Uitgeverij LEMMA B.V., Utrecht

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK,
DEN HAAG

Issues

Issues in public examinations: a selection of the
proceedings of the 1990 IAEA Conference in
Maastricht, The Netherlands/ed. by
Anton J.M. Luijten. - Utrecht: Lemma
ISBN 90-5189-091-5
NUGI 724
Trefw.: examens; opstellen.

ISBN 90-5189-091-5
© by CITO and by the authors

UITGEVERIJ LEMMA B.V., Postbus 3320, 3502 GH UTRECHT

All rights reserved. Except for the quotation of short passages for the purposes of criticism and review, no part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the publisher.

Cover: Andries Harshagen, Culemborg.

Preface

The 18 chapters in this volume constitute a selection from the 57 papers presented at the 16th Conference of the International Association for Educational Assessment (IAEA) held in Maastricht, The Netherlands, in June 1990. This collection reflects the attempts by examining bodies and institutes to improve examination systems and examining techniques, to develop reliable instruments and to establish standards in public examinations.

These chapters represent a great diversity of professional interests in educational outcomes by means of public examinations and a variety of opinions about the value of the examination systems that prevail in various countries these days.

This publication will be of interest to those engaged in the development of public examinations but also to those involved in examination policy and interested in the examination systems and examination policy.

I like to express his thanks to the contributors for allowing papers to be included in this volume. Editorial policy has been to interfere as little as possible with each contributor's presentation as to style and terminology.

I would also like to thank my Cito colleague, Jos Holtzer, for his efforts in moulding the raw material into the beautiful publication you have before you.

The editor

Contents

Preface 5

1 Public Examinations, an Introduction
A. J. M. Luijten 9

2 Social and Educational Imperatives for Changing Examinations
P. J. Black 13

3 The Estimation of Bias in Matriculation Examination Scores and a Proposal for Correction
S. Cahan 29

4 Recording Public Examination Performance – Usefulness, Accuracy, Ownership
J. S. Cook 39

5 Divergence of Interests: Problems of Implementing Reforms in an Established Public Examination System
R. K. Crocker 45

6 Attributes and Applications of a Pro-Active Assessment System
R. Ferguson 57

7 Social Justice, Public Interest and the SAT
C. Fincher 65

8 Issues in Assessment of Non-Objective and Objective Examination Tasks
J. Izard 73

9 Determining Equivalence among European Educational Qualifications
S. Johnson 85

10 Examinations in Europe: Three major Streams
H. P. J. Kreeft 101

11 The Politics of Assessment
D. L. Ongom 111

12 Why Assess? Evaluating the Assessment of the English National Curriculum
C. Robinson 121

13 Subject Pair Comparison Studies in an International Context
J. Thompson, M. C. Hayden and I. J. Kerr 129

14 The Use of the Item Response Theory in the Dutch National Assessment Programme
N. D. Verhelst and Th. Eggen 143

15 The Rationale of a School-Based Centralized Examination System
W. H. F. W. Wijnen 149

16 Measuring Educational Progress with a Longitudinal Testing System
T. L. Willoughby and A. R. Bixby 161

17 The Wessex Project: A Modular Development in English Post-Compulsory Education
J. Wilmut and R. G. Rainbow 175

18 Experiments on General High School Completion Tests in China
Xuewei Yang 187

List of Contributors 191

1 Public Examinations, an Introduction

A.J.M. Luijten

The short story entitled *Charles Ryder's Schooldays* by Evelyn Waugh contains a very peculiar description of the purpose of an examination: "The exam is only to keep out undesirable types". The only conclusion that I can draw from this statement is that we – inasmuch as we are interested in conferences on evaluation issues and take part in discussions about the effects of examinations on education – are all desirable types.

To consider the effects of examinations on education it will be more helpful to concentrate on descriptions such as this: "In every country examinations represent the stamp of official approval set on the completion of the pupils' school career. Besides marking the end of that career, they are indispensable for inciting adolescents to a serious and sustained effort. When there is an examination ahead of them, pupils and teachers can no longer behave as though they were at home, in an easy, casual way, with allowances made for good intentions, individual temperaments, passing indispositions, changing seasons, etcetera. All must march in step and try to keep in line. Pupils pay more attention to their work and teachers make their lessons more accurate and concise. The examination must be passed at all costs." It is obvious that we are now no longer in the world of fiction. This definition is a more practical one, formulated by a worried headmaster; it is from Buisson's *Dictionnaire de Pédagogie*, published over a hundred years ago: in 1887. This definition reflects the 19th century assumption that examinations possessed an absolute value as a method of selection.

Contemporary society is discovering more and more factors tending to broaden and humanise the definition of an examination. It ought to be clear that examinations are not the be-all and end-all of education and the teaching process. They are not an end in themselves. Therefore we have to establish a connection between examinations and the ultimate purpose of education: to encourage the development of citizens who are socially, culturally and technically flexible. The ultimate goal of assessment in general is to contribute to this educational purpose. I think that in our days all of us will agree with this concept.

The theme of the 16th annual conference of the International Association for Educational Assessment held in 1990 in Maastricht, The Netherlands, was: 'Issues in Public Examinations'. This theme however was not meant as a philosophical one. The presentations delivered at this conference were dealing with educational, technical and political aspects of examinations. Experts in the field of the development of public examinations are these days confronted with the following description of a public examination: an examination is an assessment procedure (usually separated from the classroom situation and following on a learning process) requiring candidates to perform the assignments set to them or to answer questions after which they will receive – or not, depending on their

achievement – a certificate. 'Public' indicates that the examinations are taken by a large number of candidates – in case of the official school leaving examinations at the end of secondary education – in a more or less disclosed procedure formalised by the government or examining body. 'Public' also means that candidates who pass the examinations will receive a certificate which give them special rights and which is recognized by society.

The conference mentioned above focused on the following aspects:
- the comparability of examination systems and certificates used in various countries;
- developments in assessment techniques and examination procedures;
- changes in examination systems as a result of changing political, social or educational views.

The first aspect was covered by presentations dealing with the problem of comparing examination systems and certificates in various countries. The presentations showed that the attempts to establish equivalence of diplomas and certificates in an international context are still in their first stage. It also became clear that international recognition of certificates attached to examinations at the end of secondary school is hampered by the difficulty of setting down universal classifications of examinations and diplomas in view of the great diversity of subjects as well as educational levels. We have to be aware of the fact that there is a great disparity in the numbers and types of educational systems and consequently in the examination systems. Another fact is the absence of a uniform terminology or common vocabulary to describe assessment procedures and levels of achievement. Paper readers emphasised the need for greater coherence in this area and made a plea for opening the way to a better co-ordination between the various national systems; they made suggestions for solving the thorny problem of equivalences and stressed the need for international standards. During the conference two workshops were devoted to this topic. One dealt with vocational qualifications in European countries and the other with examining in an international context.

Presentations concerning the second aspect, developments in examinations procedures and examination techniques, referred to recent developments in testing- and scoring-techniques, to the new possibilities of the computer in large-scale assessment procedures, to new techniques for improving the reliability of pass/fail decisions, but also to recent developments in the use of performance-based tests in final examinations and to new attempts to make examinations reflect more accurately the demands of the curriculum. Alternatives for existing examination procedures were also discussed. Papers dealt with such questions as: should assessment specialists pay more attention to valid methods of continuous assessment building up a picture of a candidate that can be consulted at any time? Or is it more useful to improve external assessment instruments and compare, at the end of secondary school, the snapshot of the final examination with the motion picture of continuous assessment? It is clear that new developments in examination contents and techniques may also be expected to have effects on teaching and learning strategies. Are examinations being used to bring about or accelerate changes in the

curriculum or do they have, on the contrary, a conservative influence? Do examinations have a stimulating or rather a stifling effect on the teaching process? Whatever the case, constructors of tests and examinations should realize that they do indeed have a great responsibility not only for their own products but also for the effects that examinations have on education.

The third aspect, changes in public examinations in a social and educational context, was covered by presentations referring to the educational social and political backgrounds of the examination system. Papers discussed such questions as: To what extent do examinations form an integral part of the educational and school system? To what extent have research findings shown examinations to be biased in favour of particular social and cultural groups and to what extent do examinations fail to provide equality for all the candidates? How can examinations fulfil the requirements of higher education and employers?

As a matter of fact the general concern about improving education is nothing new, nor are concerns about the means by which student performance and attainment are assessed. Papers discussed the values of internal and external assessment procedures from this point of view. One point to be noted is that no general agreement exists as to whether the standard for examinations at the end of secondary education should be set by the school or institution which the pupil is leaving, or by the one which he is entering or – third possibility – by an independent 'outsider': a national committee or an external examination board.

If the main function of examinations was, as Evelyn Waugh would have it, to keep out undesirable types, we might say that the examination system would be facing a crisis. Because, if the only function of examinations was to select desirable types and to eliminate the undesirable ones, they would thus undermine the school's main function: that of guiding pupils. But let us keep in mind that Evelyn Waugh is a writer of fiction and also that the teaching process is always linked with judgements and decisions about pupils' achievements, about what is true and what is false, about who is capable and who is not. Final measurements of achievement are used for selection purposes, but they also have the function of providing pupils, parents, teachers and employers with reliable information and useful guidance. I believe that we in our educational situation, with our variable concepts of what an examination is or has to be, at least agree that examinations have in view to bring out the best of the individual on both sides: the examinees and the examiners.

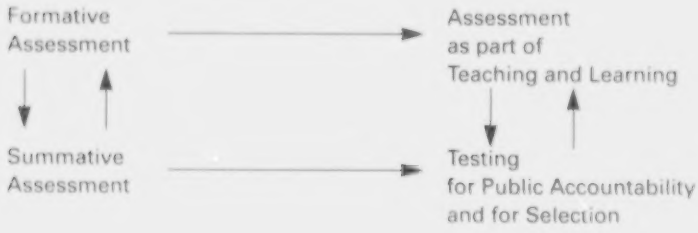
The main purpose of this conference, therefore, was to consider and reconsider the facts and findings of the different examination systems, to investigate the different functions of our examination systems, the reliability of decisions based on the examination results and to look at the strengths but also at the weaknesses of the examination procedures followed in various countries. The papers included in this book present a comprehensive picture of the issues mentioned above.

2 Social and Educational Imperatives for Changing Examinations

P. J. Black

Introduction

The main concern of this chapter is illustrated by the following diagram:



In order to explore the relationships indicated, I shall start by discussing some examples of assessment items to illustrate the functioning of assessment tasks and their relation to aims and models for pupils' learning. This will lead to arguments about the development of formative assessment with particular reference to assessment in the new National Curriculum in England. Then, by contrast, I shall discuss public examinations and the social pressures for accountability. Finally, I shall consider the tensions between the two functions of assessment and the prospect of resolving these tensions.

Examples

The three examples given below are all taken from national surveys in England and Wales undertaken by the science teams of the government's Assessment of Performance Unit (Russell, T., et al., 1988, Schofield, B., et al., 1988, Archenhold, W. F., et al., 1988, and Johnson, S., 1988). They are all concerned with science, but I believe that they illustrate points of general applicability. The surveys in which they were used pre-date the National Curriculum for Science; however, they can be related to the criteria contained in that curriculum, and they are of particular interest because they assess aims which were quite uncommon when the tasks first appeared but which now have to be implemented in all schools.

The first, shown in figure 1, is one of a group of questions which required pupils to see and to express the pattern in tabulated data. The success rate of pupils on such questions showed quite wide variations

Figure 1 Country energy – a question about interpreting presented information

Records are kept of the total amount of energy a country uses each year (from oil, coal, etc.). From this the average amount used by each person in that country can be worked out.

The table below shows the percentage of the population working in agriculture and the amount of energy used per person each year for six different countries.

Country	% of population working in agriculture	amount of energy used per person (units of energy per year)
Ceylon	50	0.8
Cuba	42	5.1
France	26	19.5
Italy	31	8.4
USA	12	66.0
W. Germany	23	26.8

Describe what the table shows about the way the percentage of people working in agriculture relates to the amount of energy used per person in a country.

between the questions. This particular question is more difficult than many similar ones, and comparisons indicate that it would be more easy if the data had been presented in numerical order, and that the inverse relationship which it displays is more difficult for pupils to describe than a direct relationship. Similar questions have been set displaying three columns of numerical data with a systematic relationship between only two of them – this variation can lower the success rate even further. This illustrates a general problem which is important in any criterion referenced system, the problem of ambiguity of criteria, which faces those who have to construct and mark assessment instruments. To develop this point, consider the three following criteria which appear as part of the specification of the new National Curriculum (DES, 1989) and which are intended to serve as criteria for assessment:

- *Level two* – interpret findings by associating one factor with another, for example, *the pupils' perception at this level that 'light objects float', 'thin wood is bendy';*
- *Level three* – interpret observations in terms of a generalised statement, for example, *the greater the suspended weight, the lower the spring;*
- *Level five* – make written statements of the patterns derived from the data obtained from various sources.

The Levels are meant to relate the criteria to the expected attainment for average pupils at various ages – age seven for Level Two, age nine for Level Three and age thirteen for Level Five. Thus they ought to specify the progress that pupils are expected to make (DES, 1988).

In practice, each of these statements has to be operationalised in terms of specific questions: the Country Energy question above could be used for Level Five assessment, but the easier variant in which the data are presented in numerical order, or the harder one in which a third column of data is added, would match just as well. Furthermore, if an exercise of this type were to be encountered by pupils in the course of tackling a larger task using their own data, then performance might be different again. This last reservation should be borne in mind in any appraisal of tasks set in arbitrary or isolated contexts.

It would be easy to produce many more examples to show that these criterion statements are not sufficiently specific. However, it is also clear that they could not be made unambiguous without making them either unintelligible, or perhaps elaborate to the point that they could only be met by one particular question – which every pupil would then learn by heart. Those constructing assessment items might resolve some of the ambiguities by reference to the age norms specified for the various levels, and if assessment items were to come from an authoritative source, they might complement the criterion statements and reduce the ambiguities.

The answers given by pupils are also relevant here. When the question above was set to 15-year olds, 49% were able to explain the correct relationship, although only 14% were able to do this in precise terms, i.e. using explicit reference to *'percentage* of people' and to *'energy* per person'. This indicates that easier versions might have to be used for a Level Five assessment (which should correspond to success by the median 13-year old). However, this question was set in an open-ended form and any 'result' is dependent on the marking criteria which have to be fashioned to give valid categories for the range of ways in which pupils might respond. Analysis of responses from several hundred pupils to the above question showed that the following (real) examples were typical of the main types of response:

1 The less the percentage of people working in agriculture, the more energy will be used.

2 Ceylon uses the most amount of energy working in agriculture.

3 Because the people working in agriculture use more fuel than the average person.

4 The countries where a lot of people work in agriculture can make their own energy. In industrialised places they would have energy running into their homes off a main energy giver.

The first of these is a good answer – perfect except for the omission of 'per person' after 'energy'. The second is a common response to questions of this type. The pupil selects an extreme value and states it: it is not possible to say whether such a pupil has misunderstood the demand, or is unable to meet it whilst having a glimmering of the relationship through looking at extreme cases, or can perceive it but is unable to express it. The third and fourth answers represent another very common type of response the pupil

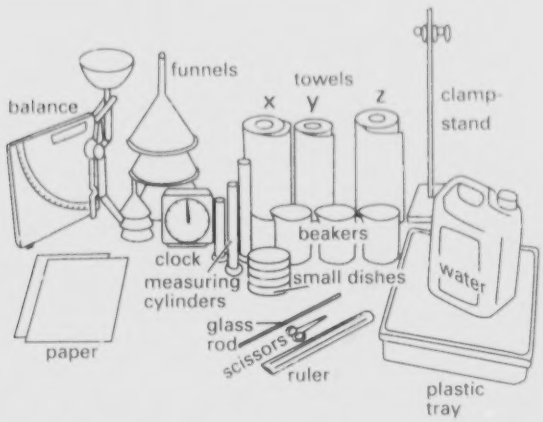
attempts an explanation – because the demand of the question is not understood, or because the pupil's experience has always been that science questions demand explanations, or because the pupil has to think in terms of his or her model of the particular context and cannot abstract the 'pattern' aspect from this context. The third response proposes an explanation which does not fit the data, but the fourth is a valid and ingenious attempt; it is ironic that this able pupil must be recorded as failed in company with the third.

Here the example illustrates the value for formative assessment of sample tasks which have been extensively used so that a representative range of pupil responses is available. Such examples can indicate the potential of a question to evoke pupil response and can help teachers to foresee some of the difficulties with which pupils may need help. In addition, the ways in which the success and character of responses change as the nature of the questions about a particular skill or topic is altered, can give useful empirical evidence. Such evidence can help to identify those features which characterise progression in the particular aspect of pupil performance which the questions reflect. The development of formative assessment, and of its related criterion statements, will require that criteria and practice are refined by empirical evidence in this way.

Figure 2 Paper Towels – an open investigation task presented to pupils with equipment

You have in front of you three kinds of chapter towels labelled X, Y and Z
This is what you have to find out:

Which kind of chapter will hold the most water?



You can use any of the things in front of you. Choose whatever you need to answer the question.

Make a clear record of your results so that I can understand what you have found out.

The second example (Gott, R., Murphy, P., 1987) is a task which was first used as a practical task, in which pupils were presented with the specification and the equipment illustrated in figure 2.

Performance was assessed using observation of each pupil individually. The first surprise was that many pupils tried out the various chapter towels by soaking up water with them and came to a decision to select the best on their qualitative impressions. These same pupils were able to use the measuring instruments provided when directed to do so with a specified task, but they did not choose to use measurement when faced with this problem. Other pupils taught us about the difficulty of the task through the variety of ways they found to interpret the phrase 'hold the most water'. Some tried to make a small cup out of a piece of chapter towel and then poured water into this cup, some put a piece of towel into a measuring cylinder containing water and then measured the rise in the level of the water, some put a piece of chapter towel in water, then hung it out to dry (like clothes on a clothes line) and measured how long it took to dry. Thus the task raised two issues about the measurement demand – whether pupils knew when to attempt quantitative measurement and whether they could decide what to measure i.e. transform the demand in the question into

Figure 3 Paper Towels – comparison of pupil performance in planning or performing investigations

Overall performance level	Description	% pupils (age 13 only)		
		Prose (n=501)	Pictorial (n=398	Practical (n=824)
1	An accurate quantitative measurement method	12	24	43
2	A rather less accurate method but still capable of discriminating between the towels	3	8	5
3	Either 1 or 2 but with no restriction on size of towel used	13	14	19
4	Here, only a final measurement was made in those approaches where this is not sufficient	1	3	10
5	Qualitative measurements only or no measurement either taken in the practical or mentioned in the written versions	71	51	23

(Data for age 15 pupils follow a similar pattern)

a set of operations to lead to measurable quantities.

A further lesson learnt from this task is illustrated in figure 3. This task was presented to three matched samples of pupils in three different ways. In the first two, the pupils were asked to give their plan for the investigation, but in the first the equipment was listed in prose only, whereas in the second a picture was provided. In the third way, the actual equipment was given and pupils performed the experiment. Plans and performance were assessed by the same criteria. The results showed clear – and very large – differences between all three. These results show that if we were to rely on chapter and pencil methods to give evidence of pupils ability to carry out investigations we would be seriously misled. They also show, between the first two types, another effect of the mode of presentation.

My third example is a pair of practical tasks presented to pupils with equipment with individual observation of each pupil's actions. figure 4 shows the questions (Schofield, B., et al., 1988). They were designed to be

Figure 4 A matched pair of investigation questions

SWEETS

When you eat certain sweets they dissolve in your mouth and fizz. They do the same thing in water.

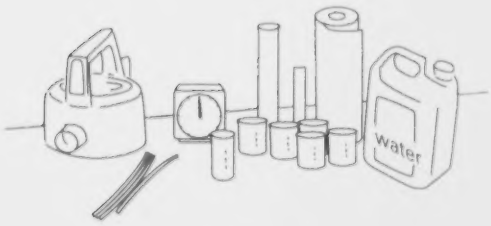
This is what you have to find out:

What makes the difference to how long the sweets last for?
Is it how hot or cold the water is or how much the sweet is broken up?

You can use any of the things in front of you. Choose whatever you need to answer the question.

a Make a clear record of your results so that someone else can understand what you have found out

b Write your answer here.



identical in respect of the structure of the tasks and of the demands that they entailed, except that the first (Sweets) was set as an everyday problem, expressed in everyday language and provided with kitchen equipment, whereas the second (Chemical) was set as a science problem with scientific terms and laboratory equipment. Each was set to a different sample of about 250 pupils, the two samples being matched on those criteria which were known to affect performance. The performances of the two groups were then compared on a number of criteria.

Figure 5 shows the results for 'Manner of Measurement'. The pattern it shows is typical of all such results: the average performances were better for the version set in a science context than for the everyday context. This is but one of many examples showing that the way in which the demand of a question is presented can affect response. Whereas this example was concerned with the effect of context, other features, such as the language used, whether the information is presented with pictures, with prose, or the

CHEMICAL

When certain chemicals are placed in water they fizz and bubbles of gas are formed.

This is what you have to find out:

What makes the difference to how long the chemicals last for?
Is it how hot or cold the water is or how much the chemical is broken up?

You can use any of the things in front of you. Choose whatever you need to answer the question.

a Make a clear record of your results so that someone else can understand what you have found out

b Write your answer here.

Figure 5 Comparison of results: manner of measurement for sweets and chemical (each figure is the percentage of pupils meeting the defined criteria)

Measurement stage-level	Description	% pupils	
		Sweets	Chemical
1	Timed until all dissolved, used the clock accurately	26	54
2	Timed accurately but failed to use an accurate end point of measurement	24	23
3	Clock not used accurately, end point judgement inaccurate	24	9
4	Qualitative judgement of end point	24	14
	No apparent judgement	2	–
Number of pupils		257	248

layout of the question all affect success rates.

A further lesson arises by comparing the Paper Towels task with the Chemical task. Those formulating the National Curriculum proposed to include a section on the 'Exploration of Science' and included in this criterion statements about various component skills e.g. the statements quoted in relation to the first example in this section. They also wished to specify that pupils should carry out their own investigations, and so had to formulate criteria to guide teaching and assessment and to describe progression in such work. Here are three of the criterion statements which show part of the set within which this formulation was achieved:

- *Level two* – ask questions and suggest ideas of the 'how', 'why' and 'what will happen if' variety;
- *Level four* – plan an investigation where the plan indicates that the relevant variables have been identified and others controlled;
- *Level seven* – plan, design and safely carry out an entire investigation drawing on their scientific understanding at this level, to test a hypothesis in which they identify and manipulate two continuous independent variables and control others (other detailed criteria follow).

It is clear that a pupil attempting the Paper Towels task would have the opportunity to show that he or she could meet the Level Four criterion but not the Level Seven criteria (DES, 1988). A pupil could of course show that he or she was only competent at a lower level, for example, by failing to identify the dependent variable. Thus the task could indicate the stage of progression through differentiation by outcome. However, the range over

which it could do this is limited, since it gives no opportunity to 'draw on scientific understanding' apart from understanding of procedures and it hardly qualifies as embodying scientific understanding in a test of a hypothesis. Furthermore, the task presents, as the independent variable, three distinct entities – the three types of Paper Towels – so that it cannot qualify for the criteria of Level Seven which specify continuous independent variables. The Chemical task however, does have two continuous independent variables, the temperature and the degree to which the chemical is broken up, although it might still fall short in testing of a hypothesis.

These features illustrate that progression may relate to the nature of the task with which a pupil is presented, particularly to the variable structure inherent in it and to the understanding of science concepts required to pursue it effectively. The measurement demands mentioned above for Paper Towels could be another aspect in which different tasks might be suitable for different stages of progression. Thus progression in this area requires differentiation by task. In practice, this understanding of open investigations was first developed mainly by those composing assessment tasks and studying pupils' responses to them in the A.P.U. monitoring surveys. However, as the ideas have been developed in the form of criteria for progression, illustrated by tasks such as those shown above, and by the range of pupils' responses, what has emerged are guides to learning and teaching, the criteria and the assessment being secondary to, and supportive of, the learning aims.

Formative assessment

The examples above illustrate several lessons about assessment in general and formative assessment in particular. Country Energy illustrates:

- the inevitable ambiguities in translating criterion statements into assessment items;
- the variety of lessons that can be learnt from the explanations that pupils were required to provide;
- the value of these for the formative function of this as an assessment task and, in particular, the clues they can provide about progression in learning;
- the problems that would arise if this task had to serve a summative function, both in respect of criterion ambiguity and in respect of the possibility that some pupils would give incorrect responses because they misunderstood the demand.

Paper Towels illustrates:

- the potential for differentiation by outcome in an open task;
- that written tests cannot be a surrogate for practical performance tests, which is as much a lesson about learning as about assessment;
- that study of the responses of pupils can teach us lessons about the demands inherent in a task and so can expose learning needs which may have been neglected.

Chemical and Sweets illustrate:

- the large effect on performance that follows from changing the context of a task, which is again a lesson about learning aims and progression as well as about assessment methods;
- the identification of progression in such learning through differentiation by task, linked to analysis of the demands inherent in a task, which relates, in a science investigation, with the variable structure and the measurement demand.

Four more general lessons follow from this summary. The first is that assessment tasks which are rich in lessons about pupils' learning may not be suitable for summative assessment unless the very diverse responses of pupils, that is part of this richness, can be interpreted in the light of other performances within a collection of such tasks, or from a wider knowledge of the individual pupils concerned.

Secondly, the lessons can only serve as a guide to pupils' learning needs if pupils' responses can be interpreted. Such interpretation can only be achieved in the light of some model of learning which serves to generate a sequence of criteria set out to express the desired or anticipated learning progress. Furthermore, a set of tasks will not function to produce the responses that the formative use needs unless they have first been designed in the light of the criteria for progression and the model of learning which underpins them.

Thirdly, the ambiguity in matching tasks to criteria can only be resolved by exemplifying criteria by typical tasks and typical pupil responses to them, perhaps by a cluster of tasks for each criterion rather than one definitive task. One might then question whether the inevitability of ambiguity detracts from the value of the criterion referenced approach: it clearly does not if the purpose is formative assessment.

A fourth point is illustrated by the way in which the system of levels in the new national curriculum for England has been set up. The curriculum targets are specified by criterion statements arranged in a sequence of ten levels. The statements are to be so arranged that the average pupil should attain Level Two at seven years of age and progress at one level every two years. The set of ten then covers the whole of compulsory schooling from age five to age sixteen, allowing for the differences between children by having ten levels. Whilst the level statements are criteria, they can only be guides to progression in learning and to appropriate assessment targets if they are linked to expected norms for various ages. This should not be regarded as a dilution of the principle of criterion referencing; any criterion system is influenced, either in its formulation or in its interpretation, or in both, by norm considerations. It would in fact be better to regard criterion and norm referencing as two aspects of every assessment system – systems differing through the way in which these two aspects function and interact (Black, P. J., Harlen, W., Orgee, A., 1984a, and Black, P. J., Harlen, W., Orgee, A., 1984b).

In the light of this analysis, the new system in England can be seen as a very ambitious attempt to implement criterion referencing in such a way that the formative purpose of assessment can be given a central role in teaching and learning.

Summative assessment

One purpose of the discussion in the previous sections is to establish a perspective within which the typical external public examination can be appraised. Such examinations have to concentrate on performance which is measurable within the constraints of relatively short tests in which there is no personal contact between examiner and either the examinee or the examinee's teachers. In this extreme form, many aspects of performance, for example practical work in science, cannot be assessed; the use of chapter and pencil surrogates may heighten the problems by giving a dangerous illusion that they have been solved.

Even more serious are the problems associated with a commitment to exclusive use of fixed response questions. These test a very narrow range of abilities and introduce dangerous bias against pupils who do not perform well in response to their rather peculiar demands. Furthermore, such testing can cut the examiners off from any understanding of pupils thinking, about which they can, and do, make inferences on the basis of statistical data alone, a practice fraught with all the well-known dangers of inferring causes from correlations. Such examinations have low validity in that they cannot reflect many important educational aims e.g. the power to formulate an explanation. They are espoused in the name of reliability, the extreme version of this being the belief in sections of the testing community in the U.S.A. that no other approach can give a reliability adequate for public purposes. Where the criterion for reliability is consistency of responses across items, it may simply be secured by ensuring that any that do not share the bias of the majority are removed.

More generally, in this particular pursuit of reliability, the price paid in terms of loss of validity, of unhelpful feedback and of disconnection from the requirements of formative assessment, is far too high. Thus, for example, whilst coded answer questions have been included in the public examination systems in England, there has never been a serious proposal that they should completely replace other assessment techniques rather than be added alongside them as a useful addition to the range. These reservations apply with particular force to international comparison exercises. The need for comparability across countries has led to almost exclusive reliance on coded-answer questions. However, the results of these very narrow tests assume enormous significance with the public and with politicians, who do not understand that they do not reflect at all such important aims as practical work in science or the ability to express one's reasoning (Lapointe, A. E., Mead, N. A., Phillips, G. W., 1988).

External examinations attempt to reflect the whole range of a pupil's attainments within relatively short terminal tests. This summative function is mainly required by external purposes and not by internal demands of guidance for teaching and learning. However, with such summative assessment there usually arises a validity gap, between the wide range of aims in learning that can be pursued and evidenced in classrooms, and the sub-set of these which can be appraised within public testing constraints. There have been serious attempts to close this gap. Some attempts involve use of a wide range of techniques in external examinations – oral tests, practical tests, tests involving group discussions. Others have involved use

of teachers own assessments within a system which implements monitoring or moderation to secure comparability and fairness between schools. It is significant here that the degree to which procedures in the latter category have become acceptable, varies very widely between different countries; in some, no public external tests and a very light monitoring of school's own appraisals is acceptable, in others there is no prospect of school based assessment playing a part.

The choices here can only be understood in terms of the political and social differences between countries. The public perception is all-important here. Public examinations are seen as assurances of fairness and reliability to a degree that is quite unjustified. The general case, that even if the validity gap could be closed, the necessarily short external test cannot have the reliability that the cumulative results of teacher assessment could achieve, is not easily recognised. The public demand for external assessments seems to arise from three main considerations. In outlining these below, I want also to consider how those concerned to develop assessment in education might react to them. The first is the public distrust of teachers, which must be linked to the general esteem with which teachers are regarded, and to the standards of public honesty within the society as a whole. There is little that the assessment community can do about this factor apart from supporting any general movements to enhance the status of teachers. However, the more specific moves outlined below might contribute to any such movement.

The second is the belief that external examinations are technically superior and guarantee better reliability than assessments invented or administered by a teacher within the normal course of teaching. This belief goes along with a failure to appreciate that a price has to be paid, in terms of loss of validity and in terms of damaging feedback for acting on this belief by giving external examinations unqualified and uncritical support. As long as this belief is firmly held, the public – including politicians – can have little motivation to modify external systems, let alone to support any radical shift towards teachers' own assessments. Here the assessment community has a responsibility to educate opinion so that public perceptions are better informed. Few would appreciate, for example, that some of the problems raised in respect of the examples quoted above, have not even been revealed or appreciated, let alone tackled, in most public systems which produce certificates that carry high authority and esteem.

A third reason which lies behind the public demand for external examinations is the desire for assurance of common standards and of fairness of comparisons between schools. It cannot be denied that the public and politicians, insofar as they are the legitimate representatives of the public interest, have a right to expect that schools be accountable to that interest. In respect of assessment there can be two levels of requirement. These are:

- to attest the achievements and proficiency of individual pupils. It is possible to imagine a society in which this would be unnecessary but there is little evidence that the demands of either advanced or developing societies are moving in that direction.
- to appraise the performance of schools. This requirement can only be met if pupil performance evidence is considered in the light of a broad

range of evidence about the circumstances of any school – but that is an issue too large to discuss here.

It can however be said of both levels of requirement that they call for assurance of comparability of assessment standards between different teachers and different schools. They also require that the aims assessed be those that society wishes to have pursued in the education of its future citizens.

Resolving the tension

If the legitimacy of the accountability demands be accepted, then it follows that there is bound to be a tension between formative assessment, which exists only to serve the purposes of teaching and learning, and summative assessment, which exists to secure public confidence and accountability. It is important to dissolve or reduce this tension as far as possible because teachers cannot be expected to give the emphasis that they should to formative assessment if it is clear to them that their efforts are not valued, respected, or trusted by society. In addition, they cannot make the processes of assessment an important part of their dialogue with pupils and parents if it is apparent to all that these assessments will not be given public recognition and marketable value.

The dilemma is altered if it can be accepted that external assessment instruments can be of positive and direct value to teachers by serving as a model and guide for those parts of teacher assessment which they can reflect, and if they can at the same time serve as a focus for professional exchange between teachers. If teachers can play a part in administering and marking externally constructed instruments, then meetings about marking can provide a means, through the dialogue that can be built around them, for that professional exchange of concrete evidence which is the only means by which teachers can achieve comparability of their standards with one another.

Such exchange can also be part of a process whereby differences between results of external instruments and those of teachers' own assessment can be resolved, and whereby comparisons can be made between different teachers' own assessments on those aspects of pupils' performance that the external assessments cannot validly reflect. Through such procedures, it would be possible to support teachers' work in a critical and technically valuable way, and at the same time to secure the conditions for reassuring the public on the issues of fairness and comparability of standards. However, such procedures would be expensive, particularly of teachers' time, and would require extensive training, preparation, and administrative support.

Thus the overall message of this chapter rests on three main arguments – the first two of which form part of the basic philosophy for the new national assessment system in England (DES, 1988). The first is that the development of formative assessment is essential if assessment is to play its proper part in supporting teaching and learning. As a corollary to this, it

has also been argued that a curriculum definition in terms of criterion referenced statements set out to guide progression in learning is a necessary part of an effective system of formative assessment.

The second argument is that the legitimate demand for assessment methods which can secure public confidence and serve both individual certification and institutional accountability is an inescapable constraint on any assessment system. This requirement can seriously weaken or even destroy a formative assessment development unless a way can be found to relate formative and summative assessments and at the same time to secure comparability of standards.

The third argument is that the public pressures, and the acceptability of any particular solutions, are a function of wider social perceptions and expectations of schools and of teaching. It follows that those developing assessment should strive both to satisfy the needs of learning and of public confidence, and to educate public opinion about the comparative validity and reliability of the various assessment procedures that are available.

References

Archenhold, W. F., Bell J. F., Donnelly, J., Johnson, S., Welford, G., 1988, *Science at Age 15: A Review of APU Findings 1980-1984*, London, HMSO.

Black, P. J., Harlen, W., Orgee, A., 1984, *Standards of Performance: Expectations and Reality*, in: A.P.U. Occasional Paper Number 3, London, Department of Education and Science.

Black, P. J., Harlen, W., Orgee, A., 1984, Standards of Performance: *Expectations and Reality*, in: J. Curriculum Studies, 94-96.

DES, 1988, *Task Group on Assessment and Testing: A Report*, London, DES.

DES, 1989, *Science in the National Curriculum*, London, HMSO.

Gott, R., Murphy, P., 1987, *Assessing Investigations in Science. Ages 13 & 15*, in: Science Reports for Teachers, 9, C.E.S. King's College London and University of Leeds.

Johnson, S., 1988, *National Assessment: the APU Science Approach*, London, HMSO.

Lapointe, A. E., Mead, N. A., Phillips, G. W., 1988, *A World of Differences – An International Assessment of Mathematics and Science*, Princeton N.J., Educational Testing Service.

Russell, T., Black, P., Harlen, W., Johnson, S., Palacio, D., 1988, *Science at Age 11: A Review of APU Findings 1980-1984*, London, HMSO.

Schofield, B., Black, P., Bell, J. F., Johnson S., Murphy, P., Qualter, A., Russell, T., 1988, *Science at Age 13: A Review of APU Findings 1980-1984*, London, HMSO.

3 The Estimation of Bias in Matriculation Examination Scores and a Proposal for Correction[1]

S. Cahan

Introduction

In order to be entitled to the highly valued 'matriculation' diploma, Israelis have to obtain at least a passing mark (55 out of 100) on each of 7-9 subject matters. Like in the Dutch system, about half the subjects are compulsory and the remainder are optional (there are approximately 800 electives). For students who have completed high school in a matriculation-oriented track, the diploma score in each subject is the arithmetic mean, with equal weights, of two different scores expressed on the same scale – an external and an internal one. The external score is the score earned by the student on an external examination, the construction, administration and evaluation of which falls under the responsibility of the Ministry of Education and Culture. Each year, the test for each subject is uniform for the whole country and is administered on the same day nationwide.

The internal score, on the other hand, is left entirely to the individual school, within the framework of broad regulations. Note that, unlike the Dutch internal score, the Israeli one need not be based on a specific internal examination. Rather, it should reflect the students' achievements throughout high school.

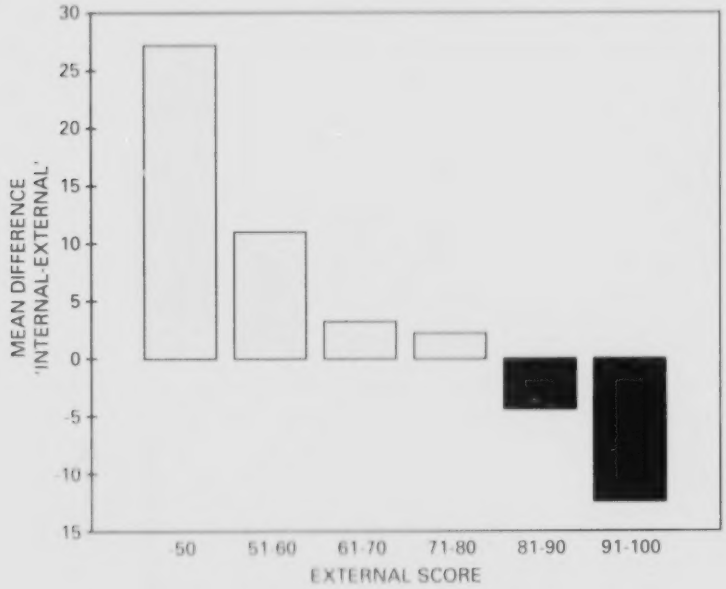
Differences between internal and external scores

Historically, the internal scores were conceived of as 'protective' scores, intended to save students from the 'merciless' external examinations. This view is no longer held today, and the rationale currently underlying this dual system is that the two scores reflect partly different, and mutually complementary, aspects of the student's achievement. Yet, the internal scores are empirically higher, on the average, than the external ones by 2-10 points (on a 100-point scale) for all subject matters. That is, they have retained their 'protective' effect. Indeed, the fact that the internal scores are the sole privilege of high school graduates is a major argument for completing high school studies. For those who drop out of high school before graduation, the matriculation diploma scores are exclusively based on the external examination.

Obviously, the difference between the internal and external scores is not uniform across examinees. Some students benefit more from the internal scores; others benefit less. Still others actually lose out, that is, their

[1] This research was supported by a grant from the Israeli Ministry of Education and Culture.

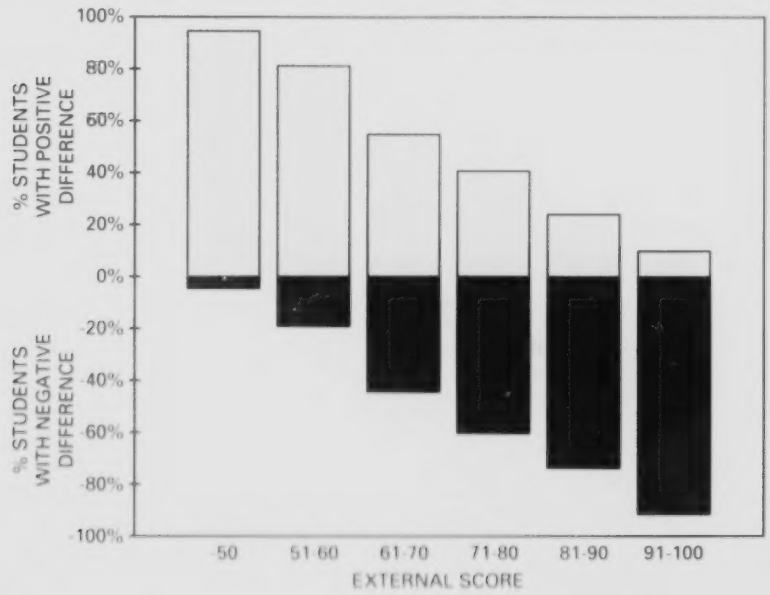
Figure 1 *The mean difference between the internal and external scores as a function of the external score – Mathematics level 1.*



internal scores are dysfunctional, lowering their diploma scores. This is true of about 25% of the students, on average across subject matters.

However, it is not the variability of differences between internal and external scores per se that is problematic, but rather that this variability is not random. There is a strong linear relationship between these differences and the external score. For all subject matters, an increase in the external score corresponds to an average decrease in the magnitude of the difference between external and internal scores. Figure 1 exemplifies this with respect to a single subject, Mathematics Level 1. As you will note, the mean difference is highest for the students at the bottom of the external score scale and decreases monotonically as the external score increases. For students scoring above 80 on the external examination in this subject, the mean difference is negative. That is, on the average, the internal scores of these students are *lower* than their external scores. Hence, internal scores play a differential role for different levels of achievement on the external examination. This is best illustrated by figure 2, which presents the percentage of positive and negative differences between internal and external scores, as a function of the latter, for Mathematics Level 1.

Figure 2 The percentage of positive and negative differences between the internal and external scores as a function of the external score – Mathematics level 1.



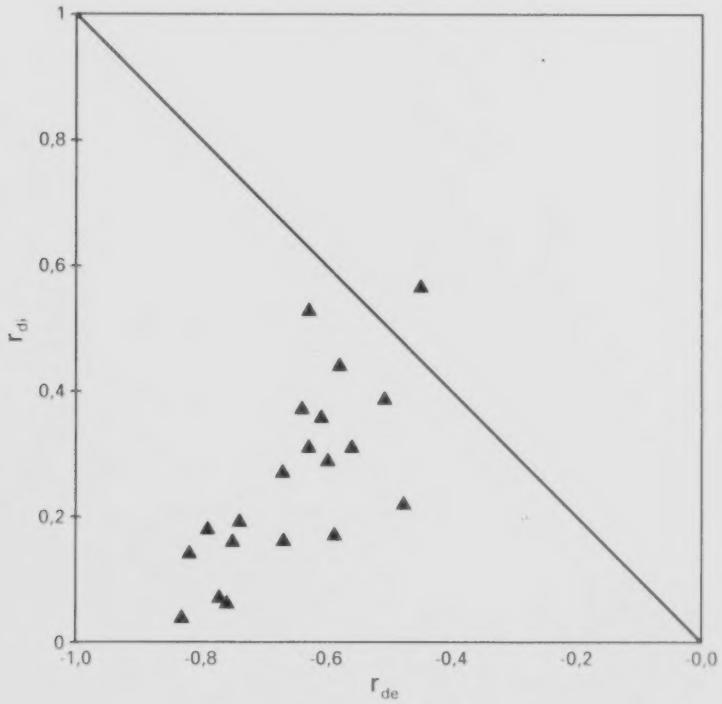
Conclusions

The results are clear and unambiguous. While most of the low achieving students (by external criteria) benefit from the internal scores, the opposite is true for the majority of high achieving students, for whom the internal scores are dysfunctional. Though they relate to a single subject, these findings are fairly representative of all subject matters.

For most subject matters, an increase of 1 point (on a 100-point scale) in the external score corresponds to an average decrease of about $\frac{1}{2}$ point in the magnitude of the difference between internal and external scores. The linear relationship between the difference and the external scores is very strong: the correlation between the two ranges from -0.60 to -0.80 across most subject matters. These high correlations clearly point to the biased nature of the difference between internal and external scores.

There are many possible explanations for this bias. The first is, of course, regression to the mean, i.e., the tendency for students with extreme external scores to have less extreme internal scores owing to the imperfect correlation between them (assuming that the ranges of the two scores are roughly equal). However, as regression to the mean is symmetric, we would also expect a linear relationship between the difference and the internal score. This relationship should be of the same strength and opposite direction as that between the difference score and the external score. This is not the case, as illustrated in figure 3: the correlations between the

Figure 3 Joint distribution of the correlations: r^{di} between the difference 'internal – external' and the internal score and r^{de} between the same difference and the external score for 21 subject-matters.



difference and the internal scores are considerably lower than those between the difference and the external scores. Hence, regression to the mean only partially accounts for the revealed bias.

A more plausible explanation is the well-documented tendency of teachers to avoid extreme scores. Indeed, the range and standard deviations of the internal scores are considerably smaller than those of the external scores. For example, while external scores range from 2 to 100 points, teachers rarely assign internal scores below 50. Thus, the biased nature of the differences between internal and external scores largely stems from the bias inherent to the internal scores themselves.

This bias negatively affects the validity of the diploma scores in each subject matter as measures of student achievement. The diploma scores of high-achieving students (by external criteria) are artifactually low and those of low-achieving students are artifactually high, resulting in an artifactual increase in the percentage of students entitled to the matriculation diploma. Moreover, the bias also diminishes the validity of the diploma scores in predicting future academic achievement. It should be noted that the mean score on an individual's matriculation diploma accounts for 50% of the scoreused to determine acceptance to Israeli universities (the other 50%

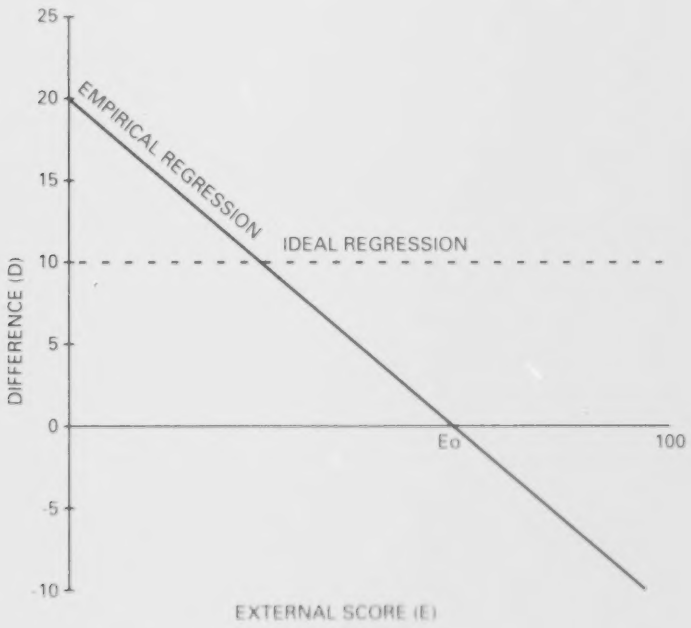
being contributed by his or her score on a psychometric test).

The problematics of the revealed bias lie not merely in its reduction of validity, but, more important, in its ethical implications. The bias introduces an element of inequity between students with high and low levels of achievement (by external criteria). For example, lower predictive validity necessarily implies lower admission chances for high-achieving students. The optimal strategy from the point of view of these students, should they be aware of this inequity, is to drop out of high school shortly before graduation, so that their diploma scores would be based only on the external examination scores.

Clearly the bias in internal scores should be corrected. While there are many administrative ways in which the bias can be reduced, its elimination requires a more drastic and direct action, namely correction of the internal scores. I will therefore now present an algorithm for the ex post factum correction of the internal scores, and then discuss some of the effects of its implementation.

The rationale underlying the suggested correction procedure is difference can be arbitrarily determined by government authorities). presented in figure 4: increase or decrease of the internal scores, such that the mean difference between them and their corresponding external scores will be uniform across external scores. (The numerical value of this difference can be arbitrarily determined by government authorities).

Figure 4 *The rationale underlying the suggested correction algorithm (K = Mean D).*



The correction algorithm is given by table 1.

Table 1 Algorithm for the correction of the internal matriculation scores

$I' = I - b(E-E_0)$

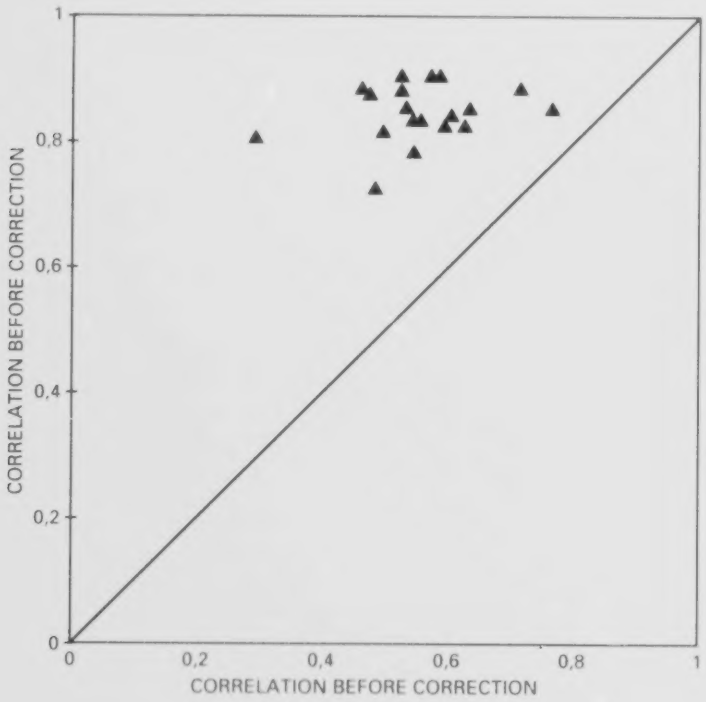
Where:

- I = Internal score
- I' = Corrected internal score
- E = External score
- E_0 = The external score for which I' = I
- b = Raw coefficient of the regression of the difference I-E on the external score E.

The correction procedure was applied to the internal scores of the 1988 cohort on 21 subject matters. Several effects of the correction are notable:

— a considerably stronger correlation between the internal and external scores (see figure 5):

Figure 5 Joint distribution of the correlation between internal and external scores for 21 subject-matters, before and after correction.



– an increase in the range of internal and diploma scores, some of which exceed 100, and a corresponding increase in their standard deviations (see figure 6).

Figure 6 Joint distribution of the standard deviations of the internal (A) and diploma (B) scores before and after correction for 21 subject matters.

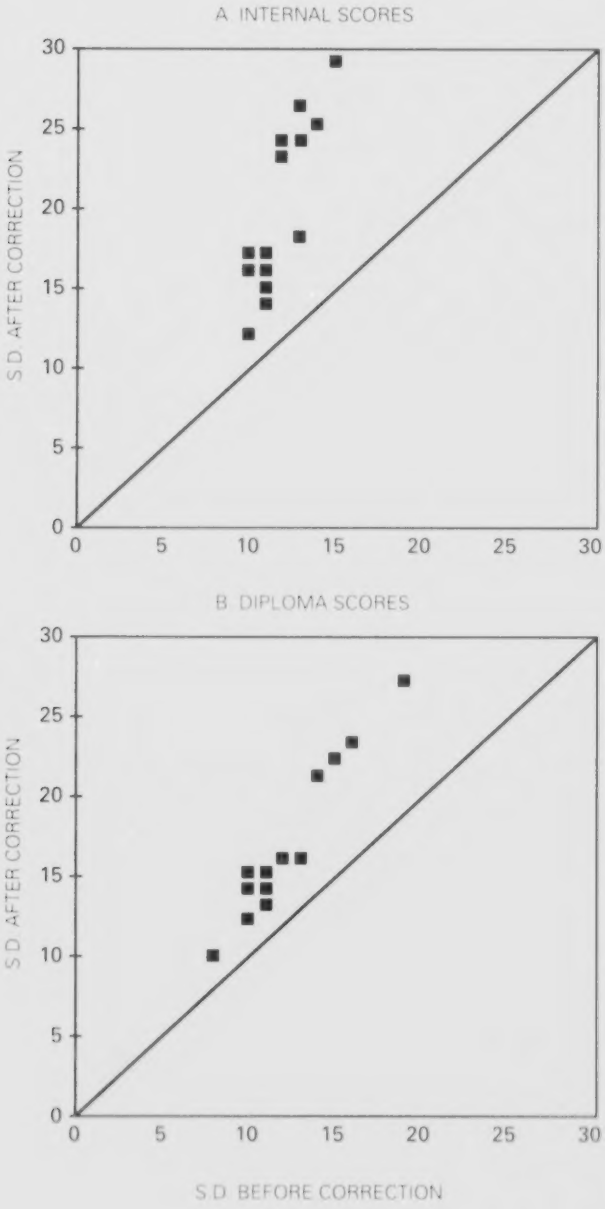
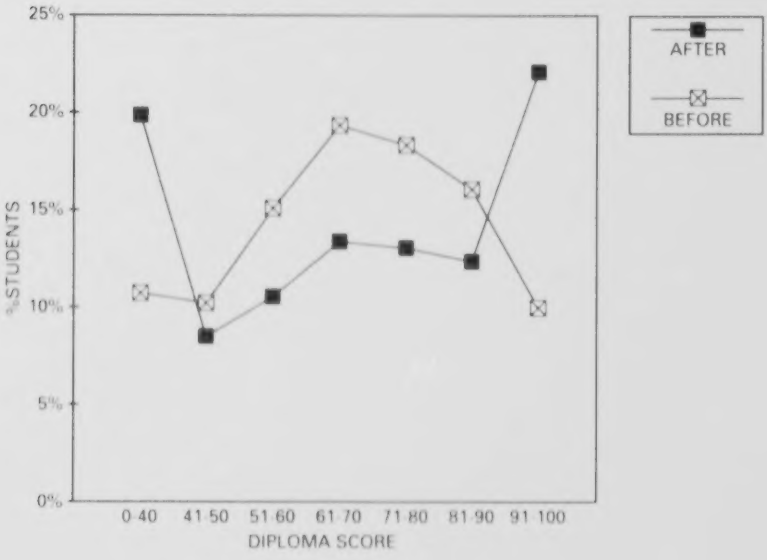


Figure 7 illustrates this effect for a single subject matter. The figure presents the frequency distribution of the diploma scores before and after correction;

Figure 7 Frequency distribution of the diploma scores for Mathematics level 1, before and after correction of the internal scores.



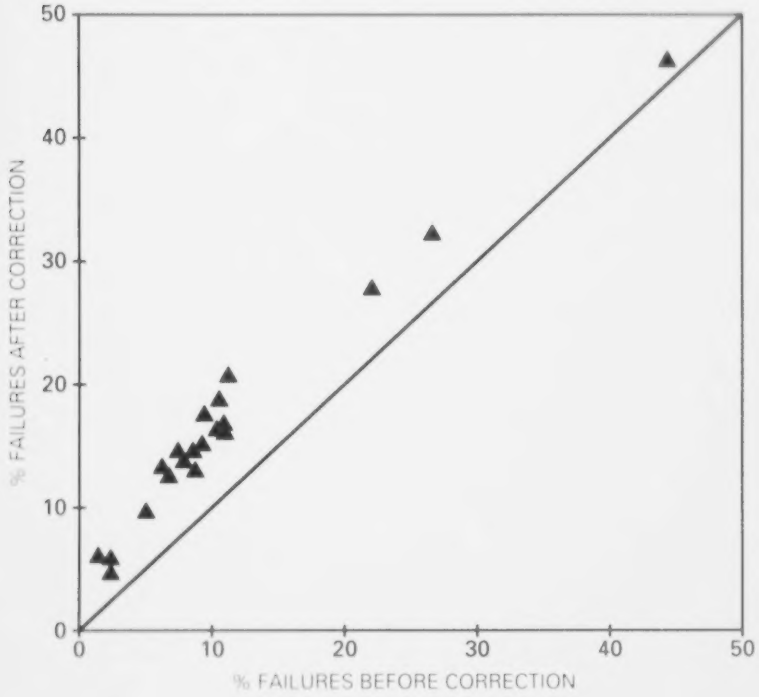
— a greater percentage of failures. (see figure 8)

Of course, a decrease in the number of those entitled to a matriculation diploma is problematic from a social and political perspective. Indeed, the Ministry of Education is strongly opposed to the suggested correction procedure, probably for this very reason. Instead, they have asked me to examine several possibilities for reducing the bias against high achievers with minimal harm to low achievers, such as:

— Consideration of the internal score only if it improves the diploma score (i.e., is higher than the external score).
— Consideration of the internal score only if the external score is above the cut-off point for failure (55).
— Consideration of the internal score only if the difference between it and the external score is smaller than an absolute magnitude, determined by government authorities.

The effects of these and other options are presently under investigation.

Figure 8 *Joint distribution of the percentages of failures on the diploma scores before and after correction for 21 subject matters.*



4 Recording Public Examination Performance – Usefulness, Accuracy, Ownership

J. S. Cook

Introduction

The reporting of students' attainments is central to the purpose of any public examination. It is an area, however, which in many systems has not received the attention that other aspects of the examination process attract. The creation of quality examination papers and the development and application of valid and reliable marking procedures are, of course, basic to any reputable examination system. These major cost items represent a bad investment if the authority responsible for the examination is not able to communicate its findings about the students undertaking the examination to various 'consumers' who wish to use the results.

This chapter examines particular issues relating to the reporting of public examination results and focuses on the often competing requirements of the total examination system. Although the chapter is based largely on observations of and experience in the New South Wales School Certificate and Higher School Certificate Examination systems, many of the issues will have an application to other systems. The chapter presents a somewhat pessimistic viewpoint in that it identifies more problems than it offers solutions and those solutions which are presented, tend to be self-evident and possibly simplistic. It may provide public examination authorities with some comfort, however, that they do have common ground even if it is the problems they share.

The public examination context

It may be worthwhile to look initially at some general features which establish the present context of public examination systems and which impact either directly or indirectly on their reporting procedures. Such features include:

- Syllabuses on which the current public examinations are based are less prescriptive than in earlier times.
- Consequently, the examination papers often have a wide question choice.
- At present, a public examination can involve students undertaking practical and oral/aural tests, and submitting projects or major works for assessment, in addition to sitting for the traditional written examination paper.
- Consequently, examinations are much more comprehensive in that they seek to measure a much wider range of syllabus objectives and tend to focus more on higher order skills than on knowledge/content objectives.

- Public examinations usually have a dual purpose with both exit credential and entrance selection/placement functions. There is also an accountability function perceived by many of the general public.
- With increasing numbers of students seeking higher education, the selection or placement function has become more important as the selection has become competitive.
- There is an increasing scrutiny of and complaints about public examination results. This is in a changing context, however, where students are being given more information about their performances.
- The papers used for public examination purposes are more often 'tradition' based than developed from established measurement principles, with examiners, both developers and markers, more likely to be subject specialists than measurement experts.
- Notwithstanding this, the quality of the current public examination papers in terms of their measurement properties and the accuracy of the results has never been higher.
- There is an increasing involvement of practising teachers in the development of the examination papers.
- There is an increased use of public examination results in tandem with teacher assessments of student achievement.
- The students' performances are reported either as a composite mark or grade or as separate indices. In either case, the external exam mark is, more often than not, used also to 'moderate' the teacher assessments from different schools to bring them to a common scale.
- Recent Freedom of Information (FOI) legislation, which provides public access to government and statutory authority records, is beginning to have an impact on public examination programmes.

Some perceptions of examination marks or grades

In the context of public examinations described above, there are a number of common perceptions of marks or grades held by the various stakeholders of the system which impact directly on the reporting procedures. These perceptions are often poorly articulated and confused but nevertheless firmly held by at least some of those who use the results.

The associated but distinct notions of 'subject demand', 'test difficulty', and 'quality of the candidature' are often confused and sometimes used interchangeably by the general public, the media and even teachers. This leads inevitably to misinterpretations of examination marks or grades and provides examining authorities with a challenge to provide a readily understandable system of reporting which clearly distinguishes these concepts. Perhaps one approach is to clearly articulate the underlying assumptions and features of the reporting system, recognising that there is a danger of 'losing the audience' if the assumptions and features do not accord, at least to some extent, with the expectations and beliefs of the target audience.

In New South Wales, for example, the Board of Secondary Education has been relatively unsuccessful in explaining the system of reporting its

Higher School Certificate Examination performances using a common average mark for all subjects. Even though the Board highlights on the credential that it is invalid to make comparisons between marks awarded in different subjects, there is still the strong expectation that a particular mark in one subject should somehow represent equivalent performance to the same mark in another subject. Where it is seen that this expectation has not been realised, confidence in the marks and ultimately the total examination system is eroded. It may be that in New South Wales, the Board did not do enough with its consumer education. This demonstrates that any successful programme of consumer education of this type is likely to be slow, incremental and without major breakthroughs in the public's understanding.

It is not uncommon for examination marks or grades to be ascribed interval or even ratio scale properties by students and the general public. This perception of examination marks is quite understandable particularly when the notion is reinforced by the statistical manipulations, including the aggregation of marks, carried out by the authority responsible for the conduct of the examination. Indeed, it may well be reasonable to use the interval scale assumption as a 'model' for processing examination marks as long as it leads to conclusions that are useful and appear valid. Examining authorities need to be cautious, however, in the interpretations they place on marks or grades derived from such processes and particularly recognise that consumers may well place interpretations on the results that were not intended. Again, the approach which might best address this problem is for the examining authority to clearly articulate what processes and comparisons are valid for the marks or grades reported. In addition, where the reported indices do not have interval properties, for example percentile ranks, then this should be highlighted for the various consumers.

The faith in public examinations which characterised earlier times has certainly lessened. There has been a continued questioning of the adequacy and accuracy of public examinations and the godlike status of the 'Chief Examiner' has changed to that of a mere mortal. Yet, in spite of these factors the community at large has a confidence in the public examination which, in many systems, teacher assessment has not been able to match. Public examinations are seen to be 'fair' or maybe 'equally unfair' to all students and free from personal bias. Such perceptions may be hard to sustain given the research evidence but they are real perceptions. As a consequence, it provides examining authorities with the challenge to 'lift their game' to make reality as congruent as possible with these perceptions. While this may be more easily said than done, the challenge remains.

Some principles for reporting examination performance

Fundamental to this challenge is the provision of information about students based on their examination performance which is consistent with a number of principles including:

— recognising the inevitable fallibility of examination marks or grades and accounting for such measurement error in reporting examination performances;

- providing information that is useful to the various major consumers. In this regard, it is essential that whatever is reported must relate to a frame of reference that consumers have firmly established in their minds;
- providing detailed information that minimizes the potential for misinterpretation by the users;
- pitching the information and the supporting explanatory materials at the level of understanding of the major user or users;
- using statistical procedures which give results that are useful and appear to be valid. (In this regard, psychometric 'purity' is desirable but not necessarily mandatory);
- not accepting, however, a situation where the validity of the results stands or falls on the acceptance of some statistical assumption or technique;
- when problems in reporting or interpretation do arise, resisting the temptation of handing over the particular function which is the source of the problem to an outside agency. (The outside agency may not do a better job, you may never get responsibility for the function back, and the changed responsibility could well impact on the total system);
- considering the possibility that the present public examinations and their associated credentials may not continue to satisfy the needs of consumers in the context of changing candidature.

It is a relatively easy task for examining authorities to recognise the fallibility of any set of examination marks. It is much more difficult to properly account for such measurement error in reporting the examination performances. The common approach of using coarse gradings derived from marks awarded by examiners, (e.g. A-E gradings based on marks on a 0-100 scale), has a number of problems associated with it. First, the inevitable misclassification of candidates that occurs in any reporting system will be more traumatic where a coarse grading approach is used. Indeed, the fewer the points of the scale, the more traumatic the misclassification will be, with the most serious situation occurring when a two point or 'Pass-Fail' scale is used. Second, irrespective of the reliability of the original test scores, the reliability of the coarse grades will be less. That is to say, rather than accounting for measurement error by using a less fine scale, the use of coarse grades actually contributes to the errors. Furthermore, the use of a coarse grading system may render the interval scale assumptions less valid and increase the potential for misinterpretation. There is also a very real danger that the bottom grades will be perceived by at least some of the target audience as a 'Fail' grading.

The reporting of examination performance as a band of marks is one approach which seeks to account for measurement error in the scores. In this method the midpoint of the band usually indicates the obtained score with the width of the band being a function of the standard error of measurement of the examination scores. Such an approach may salve the test developer's 'measurement conscience', but it is little use to the consumer wishing to make decisions based on the examination performance and lends itself readily to misinterpretation. With the focus on the error component it is a simple matter for uninformed students to

conclude they must be at the top of the band.

To provide for the proper reporting of examination performance which takes into account the notion of measurement error and is useful to various consumers, it is necessary to indicate:

- the most probable score without implying an unjustified level of precision;
- the extent of the possible error of measurement.

The most appropriate scale for reporting the performance will be dependent on the purpose(s) of the examination and the test instruments used. It may be desirable that the consumers' consideration of measurement error should focus on the interpretation of the reported marks or grades rather than incorporating the notion of error into the marks themselves.

One frame of reference that almost all consumers believe they understand is marks. A 'smart' examining authority will therefore continue to use marks in reporting examination performance ensuring that it 'encodes' them before release so that when the consumers bring their predictable frame of reference to bear in interpreting the marks, they will come up with the intended conclusions.

Competing requirements

Often there are competing considerations and requirements in public examination systems which impact directly on the reporting of students' performances. Some of these requirements are addressed in the following questions:

- Is it possible to both adequately describe students' performance on the examination and provide useful information to make selection or placement decisions about them?
- In reporting students' attainments, is it possible to be both comprehensive and comprehensible?
- Is a consumer demand for simplicity in reporting student performance possible if accuracy of the information is a requirement?
- Is it possible to provide full and accurate information about all students' performances while still having regard to the self esteem of the lowly ranked candidate?

In recent times there have been a number of alternatives proposed to the traditional methods of reporting student achievement. These alternatives, which include standards-based assessment, work-required assessment, profiling, etc, are an attempt by well intentioned educators to provide a more complete, less summary judgemental description of what students have achieved. Although such methods have a proper place in individual schools' assessment programmes, they 'fall at the last hurdle' in the public examination context because they do not meet the needs and expectations of their audiences. That is to say, they provide a mass of information which consumers lack an experience-based frame of reference to interpret. If an examining authority were to tell its consumers that a particular candidate

completed 'p' papers characterised by attributes a, b, c... and the consumer has no immediate way of knowing that p lots of a, b, c... is outstanding for the particular student population or average or extremely poor, then the consumer has been told lots but informed little. It would appear that in using such methods we keep our own hands clean at the cost of keeping the consumers' heads empty of useful but possibly harsh information. The lesson to be learnt is a simple one – you must report against a frame of reference that the consumers have firmly in their heads, or think they have. One further comment should be made regarding these competing requirements in public examination systems. If there is a situation where the information about the students, which is understood by the consumer, is less than the total information available from the examination, it is better to restrict the information provided to consumers to that, which is understood. Otherwise, there is the possibility that what is not understood, but available, will serve to cast doubt on the veracity of what is available and understood.

Marketing the 'product'

Examining authorities have never been particularly good at marketing their 'product'. The traditional focus on devising sound testing instruments and developing valid and reliable marking procedures ensures a quality product but more is needed. Fundamental to the proper marketing of the public examination system is the provision of an accurate and readily understandable system of reporting students' performance. The information reported has to be not only psychometrically sound, but able to be readily interpreted and accepted by consumers. To achieve this, examining authorities might consider using some of the marketing practices commonly used in the commercial world. It would seem reasonable that, in developing and implementing a system of reporting examination performance, the responsible examining authority could:

- identify the target audiences or consumers;
- establish the characteristics and the expectations and needs of each consumer;
- tailor the product to meet the needs and expectations of the consumer;
- if necessary, create a need for the product;
- monitor the use and acceptance of the product;
- if necessary, modify the product to meet the changing needs and expectations of the consumers.

Different examining authorities are likely to perceive the relevance of these commercial marketing practices in differing degrees and the reporting systems that might be developed will be quite different. If the reporting processes are to be effective, however, they are likely to have two common features; more attention will be given to the issues of reporting than at present and there will be recognition of the need of a greater input of resources into the reporting aspects of the system.

5 Divergence of Interests: Problems of Implementing Reforms in an Established Public Examination System

R. K. Crocker

Introduction

This chapter describes a pilot project designed to investigate the feasibility of using item banks as a basis for test development in a public examination system. The project was conducted in the context of an established public examination system, which has enjoyed substantial public and professional support, but which has undergone no substantive change in many years. In measurement terms, the system may be criticized as having no procedures for establishing test validity or reliability, for being unclear as to whether results are to be interpreted in norm-referenced or domain-referenced terms, and for not making use of available technology for data handling and analysis.

The item banking project was intended to overcome some of the measurement problems of the system, while retaining its basic goals. From the outset, item banking was viewed as a tool for improving test validity and stability. At the same time, the item banking project was embedded in a broader agenda for change, which included shifting the balance of resources from manual marking to test development, establishing performance criteria, and improving the efficiency of data processing, record keeping, and reporting.

Background and context

Education in Canada is a provincial responsibility. Each of the ten provinces thus operates independently in developing its own curriculum and assessment system. Until recently, there was little interaction among the bodies concerned with assessment in the various provinces, and it is possible to find as many different sets of practices as there are provinces. In recent years, there has been much more interest in education as a matter of national concern, and particularly in the issue of common national indicators of achievement. Provincial groups have become more aware of each other's work, and a small 'invisible college' of measurement and assessment specialists can now be said to exist. A further significant development is that several provinces have reinstated public examinations after nearly two decades of relying on school assessments as the mean of certifying students for secondary school graduation This move has clearly come about as a result of public concern over a perceived decline in standards of attainment.

Newfoundland is one of only two provinces (the other is Quebec) which retained a full public examinations system throughout the 1970's and early

1980's. Since the early 1970's, students' grades in final year high school courses have been a 50/50 blend of school and public examination marks (a system referred to as shared evaluation). In order to maintain some consistency in school marks, the latter have been moderated using provincial norms.

Since the introduction of a credit system in the early 1980's, high school certification has been based on credits accumulated over a three year programme, even though only the final course in a subject area is examined through public examinations. The overall credit system is maintained through a central students records system which, until the item banking project, had been the only part of the system making use of a computerized data base.

Local concerns with achievement standards, combined with the moves to restore public examinations in other provinces, have drawn public and professional attention to the system in recent years. Some have used the fact that other provinces are returning to public examinations as an argument supporting the wisdom of having retained the system throughout a period of substantial pressure for its abandonment. On the other hand, there have been strong indications that the system has led to a perpetuation of low achievement standards, and to criticisms that the system has not kept pace with developments in curriculum, in measurement theory, or in data processing capacity. In contrast, some other provinces have essentially been able to rebuild their systems, capitalizing on new technology and emerging theoretical and practical developments in assessment.

In Newfoundland, all final-year high school subjects have public examinations.[1] Since a revision of the secondary programme several years ago, examinations have had to be prepared in more than 30 different courses. Because of the small population of the province, some of these courses have quite low enrollments, in some cases fewer than 500 students. The unit costs of maintaining such a system could therefore be quite high if normal test development practices were followed.

In practice, the problem of cost has not arisen because the public examinations consist essentially of teacher-made tests, which are not subjected to the usual validity and reliability checks. Each test is prepared by a single 'setter', who is paid a fee equivalent to a couple of days' work. The only check against the work of the setter is a review of the test by a provincial curriculum consultant. In practice, a pattern emerges over time, with each year's test essentially evolving from those used in the past. Most tests consist of objective items comprising 30 percent of the total mark, and written response items comprising the remaining 70 percent. Because of these proportions, a large scale manual marking effort is required. (Objective as well as written response items have been manually marked). About 300 teachers are employed for a couple of weeks in July of each year, marking a total of some 75,000 papers.

Since the system has remained relatively small by the standards of large scale testing programmes, it has not been necessary to rely solely on

[1] In contrast, the provinces wich have recently restored public examinations, have done so only in selected core subjects. On the other hand, Quebec examines far more subjects even than Newfoundland.

multiple choice items. On the surface, it might appear that a system which remains small enough to employ manual marking would allow highly sophisticated items to be used, testing the highest level cognitive processes. In practice, however, this is not what has occurred. An early analysis leading up to the item banking project, revealed that most items have been constructed and scored in such a way that only relatively low level processes are involved. For example, the quality of a constructed response is rarely considered in marking. In terms of what was required of students, most items in the tests analysed could easily have been cast into multiple choice form. Thus, if the main question was one of efficiency, there was little doubt that much of the cost of marking could be saved by a simple conversion of the tests to a mainly multiple choice format. However, since the intent was to improve quality under circumstances in which it was unlikely that additional resources could be found, an important question was whether savings in marking costs could be used to achieve the desirable end of improving the test development process. The item banking project was thus initiated with this question in mind.

Basic procedures

The initial stages of the process were fairly straightforward. Biology was chosen as the first subject for development because it was a stable, high enrollment course (some 6000 students) and because a reasonably comprehensive curriculum guide was available, with objectives written at a level which allowed an easy transition to test items. As it happens, the structure of the guide became a major source of difficulty, as will be described subsequently. A working committee of five teachers plus a project coordinator were assigned the task of preparing specifications for the test and assembling a pool of about 500 items. Members of this committee were released from regular teaching responsibilities to carry out the work.

Initial specifications for the test were arrived at by committee consensus, guided by the stated objective of increasing the proportion of items on non-content course objectives (i.e. objectives such as science processes or technological applications of biology). Subsequently, a survey of all biology teachers in the province was conducted which yielded ratings of the relative importance of objectives and content topics, proportion of time spent on particular activities, and proportion of the test to be devoted to each objective.

Items were either written directly from the detailed course outline or drawn from available public domain documents and past public examinations. Items were classified by objective, content area, and cognitive level, and were coded by date, writer, source and other properties useful for reference. About 80 percent of the items were multiple choice. Written response items were retained for the explicit purpose of testing application, synthesis, and other higher level objectives. The entire item pool was entered into a microcomputer-based data base management system (dBASE III), with provision for subsequent entry of item analysis and related data. Several rounds of editing followed before the items were used

in an initial field trial. Prior to first use, the bank contained just over 550 items. This was expanded to 980 items following a second round of item-writing. While the developmental work was underway, information was sent to schools about the project in general and the specific changes to the biology test. For the first time, a table of specifications was also made available to teachers.

Field trials

From the beginning, a field trial of the items was considered to be an important part of the process. The main purpose of the field trials was to obtain empirical data on item properties. A secondary purpose was to test procedures for assembling test forms from the bank, and for test administration, scoring and item analysis. In addition to the usual item analysis statistics, estimates were also obtained from teachers on item and topic importance, time, coverage, and other features designed to allow items to be assessed in terms of the curriculum as taught.

For a first round of field trials, items were randomly assigned to 12 forms, and the forms randomly distributed within 50 classes, yielding approximately 100 student responses and 5 teacher responses to each item. Because it was desired to create a pilot version of the final examination within the same year as the field trials, it was necessary to carry out the field trial testing about two months before the end of the school year. It became clear on analysis of student responses that this created serious problems, the most obvious being unrealistic estimates of item difficulty. At the same time, the responses from teachers provided highly useful data for item and test validation.

Following the second round of item development, the field trial procedure was changed to one based on the use of items embedded in the main examination Specifically, 48 test forms of 15 multiple choice and one written response item each were created and included in the examination written in June, 1989. This gave field trial data from approximately 120 students on a total of more than 800 items. This procedure created a new set of logistical and communications problems, some of which will be discussed later in this chapter. Nevertheless, from a measurement perspective, it was clear that the quality of the data from this procedure was much higher than that for the first trial. In particular, the main examination itself provided a baseline against which the field trial forms could be judged. The main drawback of this procedure, aside from its complexity, was that the first examination developed from the bank had to be constructed using non-field tested items.

Validity issues

The starting point for test validation was the table of specifications. As already indicated, this table was initially developed by the working

committee on the basis of their analysis of course objectives. Following the first field trial, empirical data from teacher ratings were used to redefine the proportions in each cell of the table. This resulted in an increase in the number of items at higher cognitive levels and in non-content objectives – a move which had serious consequences in the first full administration of the revised test, as will be discussed subsequently.

Initially, it had been thought that assembly of a valid test would be a fairly simple matter of stratified random selection of items from the bank in proportions designed to fill the cells of the table. In practice, for multiple choice items, a simple random selection was sufficient, as the item population was distributed in nearly the desired proportions. Selection of written response items, on the other hand, presented more difficult problems because the population of such items was small, items were of different values, and some items represented significant departures from accustomed forms. In practice, written response items had to be customized to yield the required representation.

Since little empirical information on item properties was available prior to construction of the first version of the test, and because of the departures from conventional practices, it was not possible to estimate how students would perform on the revised test. In practice, performance on the multiple choice items was about as expected, while the written response items proved to be extremely difficult – yielding a mean item raw score of only 1.36 out of 5. This result presented some interesting problems for validity. In principle, an item is not rendered invalid simply because it is difficult. On the other hand, the high level of difficulty seemed to point to a gap between the working committee's judgements and the manner in which the course was being taught. In simplest terms, students were clearly not able to cope with items based on higher level cognitive processes and on objectives beyond simple subject-matter knowledge. The examination drew considerable criticism from teachers, with much of this criticism being based on the perceived difficulty of the test.

While this seems to indicate that the gap was between the course as it appeared in the curriculum guide and as it was taught, the ratings given by teachers in the first validation exercise suggested that teachers did emphasise those objectives which turned out to cause the major difficulties, and that teachers wished to see these reflected in the examination. This suggested the alternative hypothesis that the gap was actually between teachers' views on the importance of various objectives and the classroom practices of those teachers.

Obviously, this led to concerns about how validity should be established in such a situation. Accordingly, in the context of constructing a second test to be used as a supplementary examination, it was decided to assemble another panel of teachers to carry out further item ratings. This panel consisted of teachers who had been members of the marking board for the first examination, and who thus had a direct sense of the difficulty of the written response items. These teachers were asked for their general opinion of the items and to rate items with respect to relevance to the course and congruence with specific objectives.

This exercise seemed to reveal a strong contradiction in teacher judgements. Subjectively, teachers were strongly of the opinion that most

written response items were very difficult and of questionable validity with respect to the course as taught. On the other hand, the specific relevance and congruence ratings were quite high. Items were also rated as of much lower difficulty than the empirical results indicated. Correlations between difficulty ratings and observed difficulties were essentially zero, as were inter-rater correlations of relevance and difficulty. These results were in contrast with those for the multiple choice items, which showed strong consistencies among raters and between ratings and observed results.

All of this suggests that there exists a fundamental problem of validity. The main source of this problem seems to be the gap between espoused objectives and objectives in the course as taught. Teachers of the course appear to subscribe to the espoused objectives when asked to perform conventional validation tasks, but recognize that these are in some sense not the real objectives taught in the classroom. The usual validation exercises therefore yield misleading results, which are detectable only through an examination of item difficulty. This leads to the problem of whether the course as described in the guide or the course as taught should be used as the starting point for validation. The course in question was a long-established one, with a stable public examination which, some have argued, had become the *de facto* guide to teaching the course. By starting with the guide, and ignoring past examinations, the item banking team was effectively forcing a fundamental curriculum change without the normal developmental and communications activities which typically accompany such change. Clearly, one of the main goals of the item writing project was to improve the validity of the examinations with respect to a broader range of curriculum objectives than are being tested under the current system. On the other hand, teacher and student confidence in the system must be maintained. An argument can be made that the best way to maintain such confidence is to make as few changes as possible. As yet, there is no solution to the validity dilemma that this presents.

Technical issues

This section examines a number of problems associated with item bank storage, retrieval, and editing, along with problems of scoring and data analysis. Some of these points are relatively minor, and involve little more than the application of normal large scale data-handling techniques. In the context of the system at hand, however, some of these matters presented certain problems which may be worth documenting, if only to give a flavour of the minutiae which must be attended to in maintaining an item bank.

First, a decision had to be made on the type of data base to be used for storage and retrieval of the bank, and on the specifics of a data processing system. Prior to the item banking project, all data processing done on the public examinations had been performed by an external, quasi-government agency under contract to the Department of Education. One of the (minor) goals of the project was to increase the data processing capability within the Department and to automate more of the routine functions of the system.

At the time the project began, there was no known software uniquely designed for test item banking purposes (such systems have subsequently appeared on the market). A decision was therefore made to use a general purpose data base management system – specifically dBase III'. While this system provided excellent capabilities for item classification, sorting, and the like, its text editing and printing capabilities were severely limited (text in dBase is stored in so-called memo fields, contained in a different file from the main data base, but editable only from the data base on an item by item basis). Added to this, it proved extremely difficult initially to keep track of all the changes being made to items, so that errors kept appearing in the data base itself long after the authors or editors had assumed that they had been corrected. This problem was exacerbated by the need for security of the bank and thus for minimizing the number of hard copies available.

All of these problems were minor, however, compared to the task of assembling items into tests and field trial forms and producing camera-ready copies of these forms. Some sense of the magnitude of the task is perhaps evident from the fact that in the second field trial there were 48 versions, with a total of about 800 items, in addition to the items designated as the ones for the examination itself. All of these items had to be extracted one by one from the data base, according to selection criteria, and copied to a text processor for formatting. Graphs and diagrams were maintained separately because the data base could not handle such material (and no text scanner was available), and located as required for placement with the appropriate item.

Up to the time of the project, there had been no attempt to gather item level data in the public examinations system. Scores had been simply totalled manually and entered into the student records system. In fact, all data handling up to this point, including the reception of marks from schools, was done manually. The only part of the process which had been automated was the central records and transcript system. This system was capable of producing summary statistics on total scores but not on item performance. It was therefore necessary to introduce a procedure for capturing item data. In the case of multiple choice items, this was accomplished quite simply by the introduction of optical scanning. Written response item data were entered manually, using staff formerly employed in compiling and checking scores. Once entered, of course, the item data could be used to generate scores.

Once the data bad been captured, standard procedures were used to generate the required item and test statistics. Interestingly, it was found that the SPSS statistical package was more convenient for both scoring and data analysis than any of the special purpose scoring and item analysis packages available, largely because the reporting formats of the latter were generally unalterable.

In addition to giving the desired item statistics for entry to the item bank, this procedure was able to generate for the first time statistics that could be used to make scaling and other decisions prior to the production of final scores. For example, the system allowed poor quality items to be detected and discarded, and scores to be recalculated. In contrast, it had been common practice previously to give all students a mark for an item judged to be doubtful, or for the marking board to make adjustments based

on group judgement rather than on evidence as to the behaviour of the item.

An interesting side effect of the use of more multiple choice items and the availability of item statistics was that many more items were judged doubtful than had formerly been the case. Part of the problem was, of course, that items had not been field tested. However, it is clear that markers had a much greater tendency to question the quality of a multiple choice item than of a written response item. Typically, the queries centred around the use of specific distracters, the possibility of multiple correct responses, and other features unique to the multiple choice form. More generally, there was a concern with the absence of marker discretion in determining what to accept as correct. Item analysis information was not sufficient to resolve these problems in the minds of markers. Since the markers were given the final say on whether an item should be counted, many more items were lost in the revised test than had typically been the case in previous tests.

Standards and the interpretation of scores

Fundamentally, the public examinations are criterion-referenced tests designed to measure student achievement in specific domains of knowledge. The intent is to certify students as passing or failing or, more generally, as having attained a specific level of proficiency in the subject at hand. In practice, however, normative considerations enter the process at various stages. For example, in the absence of data on test stability, there has been no way to separate shifts in student populations from changes in test difficulty from one year to the next. Despite enrollment changes and other sources of population shift, the solution to changes in marks from year to year has been to scale to a more or less constant pass rate. In fact, the scaling process has allowed upward shifts but not downward shifts in pass rates, with the consequent possibility of grade inflation. All of this is driven by a desire to ensure that students are not disadvantaged by any possible (in practice unknown) changes in test standards. This, in turn is related to the issue of public perception of the system, the impact on student prospects for entry to further education, the careers of teachers, and a variety of other considerations which go well beyond the properties of the tests as measures. The effect of all of these considerations is that final grades are essentially treated as normative measures, even though this is not acknowledged explicitly.

One of the primary aims of the item banking project was to improve the stability of the tests or, more specifically, to provide evidence on test stability. This, in turn is related to the issue of standards. Under the current system, there is no way to respond to concerns about declining standards, because, as pointed out above, shifts in test difficulty cannot be separated from other sources of change. The item bank provides two mechanisms for dealing with standards. First, the field trial data can yield predictions as to the performance of a given population of students on particular items or on a whole test. Shifts in item performance can be interpreted as evidence of change in student population or in other features of the course as taught. Second, the item bank makes it possible to track the use of items and to

include anchor items in each year's test. Over time, these mechanisms should give the evidence needed for standards-setting and detecting possible shifts in standards.

Political considerations

From the beginning, it was clear that the Department of Education was conscious of the potential political impact of making changes in public examinations. Despite some underlying public concern over a supposed decline in standards, the public examinations had not been specifically targeted as a source of major problems. In fact, arguments over standards were more likely to take the form of criticism of the school component of evaluation, and calls for a return to a single provincial examination as the basis for high school graduation. It was also clear that teachers had a substantial stake in public examinations, despite the reduction in teacher autonomy inherent in such a system.

One of the objectives of the pilot project in biology was to help ensure that the planned changes would not meet with adverse reaction in the field. It became obvious soon after the first biology examination had been written that this objective was not likely to be achieved. The revised examination received more adverse reaction from the field than any other test in the history of the system. Questions were raised in the legislature, large numbers of letters and calls were received, and members of the marking board were highly critical of the test for being inordinately difficult. Much of the focus of the reaction was on difficulty. However, the use of different forms for the field trials also drew considerable criticism. In fact, it was clear that the only thing which prevented a full public and professional outcry was the use of scaling to bring the pass rate into line with previous years.

This raises the important question of the impact of professional and public perceptions on the way the system is structured. Here it is necessary to examine the interests of several groups, specifically the test development specialists, teachers, students, the end users of the system (primarily the institutions of higher education), and finally the public at large. Some interesting implications for the interpretation of marks, and for the concept of validity emerge from a consideration of such interests.

It is clear that the interest of both students and teachers lies primarily in having students do well on the examinations. A great deal is at stake for students, beginning with high school graduation, and extending to admission to post-secondary education and to entry-level job market positions. Teachers have similar interests in the sense that having good results on public examinations has come to be seen as the benchmark of good teaching. Even though teachers may be seen as also having an interest in maintaining standards, it may be argued that this interest is subsumed when the control of standards rests with an external examination. The clearest evidence on this point comes from interviews with marking board staff: it was found that teachers simultaneously decried a perceived decline in standards, while continuing to mark so as to ensure that there is no decline in pass rates.

The primary interest of testing specialists, on the other hand, is clearly with the psychometric properties of the tests. The usual considerations of validity, and reliability, as well as the specifics of item behaviour, were the main factors in including items in the bank and in selecting items for a test. Similarly, the use of embedded field trial items was driven by an overriding desire to obtain empirical data on item properties. Although the developers were not unaware of the possible impact of introducing change, that what was judged by the developers to be trivial shifts (such as increasing the number of multiple choice items or embedding items) was perceived as highly disruptive by teachers and students, with the accompanying suspicion that these changes were at the root of the increased test difficulty.

The views of the public are more difficult to assess. On the one hand, the concerns being expressed over student performance throughout North America are also felt on the local scene. On the other hand, the public as parents would have the same interest as students in keeping success rates high. The annual announcement by the Minister of Education about success rates on the public examinations is met with considerable interest. What is not known, is the reaction in case of a substantial shift in the pass rate, particularly in a downward direction.

More than any other groups, the various institutions of higher education make direct use of public examination results. Students are admitted to these institutions almost entirely on the basis of their performance on public examinations. The view of these institutions on the state of the system is clearly on record. The clearest indication of these views is found in the fact that the post-secondary institutions were responsible for precipitating a full-scale public inquiry on mathematics and science achievement, when it was found that success rates in their own first year programmes in these areas had suffered declines in recent years. These institutions clearly were of the belief that a decline in high school standards had occurred, and that success in high school could not be taken as a reasonable predictor of success in further education. It is not within she scope of this chapter to look at the results of this inquiry. Suffice it to say that evidence of low achievement and expectations was found, and that it is clear that public examinations results have tended to create a false impression of high attainment levels.

The essential point to be made from all of this is that there is an obvious divergence of interests among the stakeholders in public examinations, and that innovations designed to respond to one sort of problem are likely to create problems of a different sort somewhere in the system. Most important, from a measurement perspective, it cannot be assumed that changes designed to meet psychometric requirements will meet universal approval or even understanding. To the extent that flaws in validity, in particular, are built into the system in order to serve the requirements of particular groups, it becomes very difficult to bring about fundamental changes.

Conclusion

Compared to the developmental effort that goes into many large scale testing programmes, this project was a relatively modest attempt at improving the quality of a testing programme. Aside from problems with the timing of the first field trials, and the usual quota of editing and text processing difficulties, it was shown to be quite feasible to develop and keep track of a relatively large pool of items, and to gather and process the required item data.[3] Some of the problems had to do with the resources available for development and editing. Other problems were a function of software limitations and the inability to carry out software development as part of the project.

All of these difficulties were quite minor,and could easily be overcome with increased resources, staff experience, and improved software. However, there remains one fundamental problem, that of establishing validity. This, in turn, is related to differences among interest groups on what is judged to be a valid test, and on the desired outcomes of the test.

As long as validity is conceived of in terms of the constructs embodied in the broad course objectives as given in the curriculum guide, there is little problem. Each of these objectives can be reflected in a table of specifications, and consensus can be reached as to their relative importance. Even at this level, however, there is some imbalance, as one of the five objectives (that dealing with basic facts and concepts in biology) has been much more fully developed than the others. Item writers therefore had available much more specific statements of objectives in this area than in others.

More important, it is clear that the course as taught departs substantially from the course as described in the curriculum guide. It is likely that the reason for this has to do with the contrasting level of detail in the treatment of objectives in the curriculum guide, as well as with the structure of past public examinations. Furthermore, when asked specifically about the relative importance of objectives, teachers indicate that all are of substantial importance. Nevertheless, when test items are prepared which reflect noncontent objectives, the validity of these items is questioned.

Analysis of correspondence, along with interviews with marking board and validation panel members yield a hypothesis about the source of the validation problem. In brief, a test is judged invalid by particular interests if it yields pass rates lower than the customary level. In principle, of course, this is not a validity issue at all or, at most, is a matter of face rather than construct validity. Nevertheless, to the extent that there is a real gap between the course as described and as taught, there is room for genuine differences of view on what should constitute the starting point for any validation exercise.

In one sense, it may be argued that face validity is the only thing that counts, in terms of public and professional acceptance of a test. On the other hand, it is clear that both curriculum developers and teachers desire to

[3] It is worth noting that the decision was made early in the project not to use item response theory as the basis for item and test calibration. This decision was made partly on the basis of doubts wether the single latent trait assumption is appropiate for the type of test being developed, and partly on the purely pragmatic grounds that it would be virtually impossible to explain item response theory to users of the test.

see a broad range of objectives tested. The question is 'why this is apparently not translated into classroom activities which allow students to perform well on non-content objectives?' The project has not yielded a satisfactory answer to this question. To the extent that this has to do with the structure of previous examinations, it might be argued that persisting with a revision which is more broadly representative of course objectives would soon lead to changes in teaching practices which would yield improved performance on the revised test, and hence improved attainment with respect to the course as a whole. Essentially, this is what is to be done over the next couple of years. With the ability to track item performance now built into the system, it will be interesting to see whether the expected change does occur.

6 Attributes and Applications of a Pro-Active Assessment System

R. L. Ferguson

Introduction

For those concerned about and charged with the education of young people in the United States, the 1980's were a period of intense collective introspection. At the very beginning of that decade, in its report, *Three Thousand Futures: The Next 20 Years for Higher Education,* the Carnegie Council on Policy Studies in Higher Education predicted that anticipated enrollment declines would create a new academic revolution. In 1983, *A Nation at Risk,* a report issued by the National Commission on Excellence in Education, warned that "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people". Numerous other studies published throughout the 1980's identified perceived defects in virtually every aspect of U.S. education.

Although not all observers have been as critical or dramatic with their conclusions as the authors of the reports just identified, there is a shared belief that much more must be done to meet the social and economic challenges that the U.S. is destined to face in the coming decade and beyond.

The challenge

While U.S. education has always been confronted with challenges borne of the diversity of its people, most agree that the years ahead will be characterized by even more difficult challenges. Most of these challenges defy simple solutions and will require a sustained and systematic approach to resolving.

- The number of 'at-risk' students in the U.S. is on the increase. At-risk students include those who are learning to hate school, becoming dependent on drugs or alcohol, failing to acquire worthwhile employment skills, becoming despondent or suicidal, and accepting failure as a way of life. Such individuals often face the dual hardships of poverty and minority status. Many simply drop out of secondary school; relatively few go on to college.
- Overall, the high school dropout rate remains high; about one in four students fails to graduate from secondary school. For minorities, the dropout rates are higher – about 35% of Black youth and 45% of Hispanic young people leave secondary school without a diploma. The U.S. Department of Education estimates that the unemployment rate for recent dropouts is twice that of their peers who graduate.

- Costs of education at all levels are escalating rapidly. During the 1980's, the average cost for a single year of college exceeded US $10,000 for the first time. The financing of education was a topic of continuing controversy throughout the decade.
- A companion issue to those identified above involved the setting of realistic standards and the development and use of tools appropriate for measuring and affirming student learning outcomes; that is, for introducing accountability by the schools to their external publics. The process of constituting these standards has introduced a further complication – weighing the emphasis on strengthening the standards against the equally important goal of providing educational equity and access for a population increasingly characterized by racial-ethnic and socioeconomic diversity.
- Many students who enter college are ill-prepared for entry-level studies. College faculties continue to complain that many students are underprepared academically and that they must invest too much time teaching students that which they should have learned in secondary school.
- Minorities, particularly Hispanics and Blacks, continue to be underrepresented in higher education. There is little evidence to suggest that this state of affairs will change any time soon.
- There is a broad-based consensus that secondary school curricula throughout the nation needs to be refocused and strengthened to provide a better balance between requirements in basic core areas and electives. There is also agreement on a related concern, the lack of strong articulation among all the stages and levels of the U.S. system of education, from preschool through postsecondary.
- Teachers need to be better prepared and better paid.
- Business and industry is becoming increasingly vocal about what it sees to be the inadequate preparation of students to become a part of the well-trained, sophisticated workforce the nation needs to sustain a strong, competitive economy in the 1990's and beyond. The perception that vast numbers of U.S. students cannot meet the educational requirements of today's workplace, much less those of the next century, is accompanied by the belief that there are too few incentives for effort and achievement in school.

When one surveys all of the problems and ills laid at the feet of U.S. educators, it is tempting to conclude that all the answers to the nation's educational challenges lie with the schools. Many are inclined to believe to the contrary, however, that the society has unrealistic expectations of education and that it must take on a greater share of the burden of achieving its goals.

Solutions

While the realities by U.S. educators are sobering, if not gloomy, there are causes for optimism. There are countless teachers, administrators, officials,

and parents committed to assuring that the nation's young people receive education of the highest possible quality.

Nearly every sector of our society appears to have become awakened to the problems that lie before us. Faced with the revelations of the 1980's, the U.S. appears to be responding with genuine commitment to finding solutions in the 1990's. Improving education has become a national priority.

States have begun mandating completion of core coursework for high school graduation. Teacher education programmes are being strengthened. Creative partnerships are being forged between schools and business and industry. At many colleges, admission policies are being studied and modified. Organized efforts to recruit and retain underrepresented populations are on the increase. Increasingly, computers are being viewed as standard educational tools. Emphasis on foreign languages and international studies is intensifying. There is renewed and growing interest in accountability, in setting educational goals, devising appropriate programmes and strategies, and applying more sophisticated methods of assessing and evaluating outcomes.

It is not the purpose of this chapter to identify or discuss solutions to the problems identified above. We take as a given, that education will be at the heart of the nation's challenges in the next two decades and that we will measure our success in 2010 by how effectively we have affected those goals and to document that we have accomplished them. Since this is a conference on assessment, the focus of the remainder of this chapter is on assessment, specifically a new 'system' of assessments developed and disseminated by American College Testing (ACT).

ACT educational planning and assessment system

In the early 1980's, even before the reports on education referenced earlier in this chapter were issued, ACT staff began a nationwide study focused on emerging issues in education. Our purpose was to match the outcomes of these studies against the purposes of ACT's various assessment programmes, including its college admissions testing programme, and to identify needed modifications, areas where new assessments were needed, and means for making assessments more useful to educators and beneficial to students.

The results of this study, which involved several years of effort by a large number of ACT staff members, was the design and development of a comprehensive model testing system that integrates an interrelated series of measures of academic abilities with assessments of interests and study skills.

The new system is designed for sequential use across the crucial early to late teen years, as students move from eighth grade through the second year of postsecondary education. The test components of the new system address a well-defined developmental span and provide coherent, articulated information useful for a variety of purposes – student and parent counseling/guidance, course selection, and development and evaluation of instructional programmes. Results from students' participation in each

Table 1 ACT educational planning and assessment system for student/programme appraisal and development

Programme	Level	Components		
		Core Tests	Optional tests Information Element	Other
8th Grade	7-9	Reading		Bio-data
primary	8	English		Interests
		Mathematics		
		Science		
P-ACT+	9-11	Reading		Bio-data
primary	10	English		Interests
		Mathematics		Study Skills
		Science		
ACT	11-13	Reading		Bio-data
primary	11/12	English		Interests
		Mathematics		
		Science		
ASSET	13	Reading	Career Skills	Bio-data
primary	13	English		Study Skills
		Mathematics		
CPE	13	Reading		Bio-data
primary	13	English		
		Mathematics		
CAAP	13-16	Reading	Writing (Direct)	Bio-data
primary	14	English		
		Mathematics		
		Science Reasoning		
		Critical Thinking		

programme, and in the series of programmes, can be used to determine readiness for transitions and to help them take developmentally appropriate steps in educational planning and decision making and in career planning.

The system has been designed to provide substantial feedback that can be understood and applied by students and parents, and by teachers, counsellors and administrators. Another significant benefit of the longitudinal nature of the information available through use of all, or some subset of the programmes, is the resource it represents for systematic evaluation of local, state and institutional programmes.

Table 2 ACT educational planning and assessment system (secondary)

Programme	Level	School/College Uses
8th grade primary	7-9 8	Structure high school course planning for students and parents
		Monitor students' progress on academic skill development
		Provide baseline for assessing instructional outcomes in selected academic skills
P-ACT+ primary	9-11 10	Provide basis for post-high school educational and career planning for students and parents
		Provide individual and group assistance in preparing for ACT Assessment
		Monitor students' progress on academic skill development
		Provide measure for assessing instructional outcomes in selected academic skills
ACT Secondary primary	11-13 12	Facilitate student postsecondary planning and transition
		Provide a final measure for assessing instructional outcomes in selected academic skills

System components

Table 1 illustrates the system graphically while tables 2-3 describe the purposes of each component and how they are related 'developmentally'. Each system component is designed so that it can be used independently of the others. However, it is the relationships between and among the components and their potential for integrated use that give the system a unique strength.

Students enter the system at the 8th grade, when they participate in a programme called 'PLAN'. The primary purpose of this programme is to stimulate career exploration and to facilitate development of a plan for the individual student's secondary school academic programme.
In the 10th grade, students participate in P-ACT+, an every-student programme designed to improve students' planning and preparation for their education after secondary school.

Late in the 11th grade, or early in the 12th grade, students participate in the ACT Assessment Programme (AAP), the college admissions and placement programme designed for use with students making the transition from high school to college.

Upon entry to college, students participate in the ASSET (for two-year colleges) and the Collegiate Placement Examination (CPE), both of which

Table 3 ACT educational planning and assessment system (postsecondary)

Programme	Level	School/College Uses
ACT Post-secondary	1-13	Aid in the recruiting and preadmission counseling of prospective students
primary	13	
		Aid in the admission decision process
		Provide a significant portion of entry level course placement information
		Provide data base for academic advising
		Serve as a baseline for assessing student learning outcomes in the basic general education foundational skills
CPE/ASSET	13	Entry level course placement decisions
primary	13	
		Identification of remedial/developmental instruction needs of students
		Screen students for advanced placement testing
		Enhance academic advising data base
		Supplement baseline data for assessing student learning outcomes in the general education foundational skills
CAAP	13-16	Identify students' readiness for upper level coursework
primary	14	
		Enhance the effectiveness of academic advising in support of the transition from lower level to upper level coursework
		Provide a terminal point for assessing student learning outcomes in the general education foundational skills

provide detailed course placement information and facilitate a smooth transition into college studies.
At the end of the sophomore year in college, students participate in the Collegiate Assessment of Academic Proficiency (CAAP), which consists of a battery of tests that assess college-level general education skills typically acquired by that time.

System Characteristics

The system has a number of characteristics that make it a system rather than a series of unrelated testing experiences:
– *Common Purpose* – Each component is designed to support students at

a key decision point in their academic preparation and planning.

- *Common Content Base* – Each test in the system measures knowledge and skills in major curriculum areas: English, reading, mathematics and science. All the tests are constructed using developmentally articulated test specifications, a procedure that ensures the content measured by each follows a logical developmental sequence across the continuum of grades 8, 10, 11/12, and 14. The tests in each component share common item formats and follow consistent reporting procedures.
- *Common/Coordinated Score Scale* – Users of the system components that share a common score scale (P-ACT+ and the AAP) are able to track student performance over time. For components that do not share a common score scale, ACT is developing procedures for interpreting test results that are based on the content relationships across the components.
- *Common and Overlapping Reporting Services* – The research and reporting services of the system provide a longitudinal perspective on the academic performance of individual students and groups of students.

7 Social Justice, Public Interest and the SAT

C. Fincher

Introduction

The uses and consequences of standardized tests (in the United States) continue to be a minor part of more significant issues concerning social justice and the public interest in higher education. As the best known and most widely used standardized test of academic ability, the Scholastic Aptitude Test (SAT) is the frequent target of critics who believe all standardized tests to impede access and equity for students who are socially, economically, and educationally disadvantaged. Ignored in most of these criticisms is the valuable information that the SAT provides about the verbal and mathematical abilities of students completing secondary education and entering U.S. colleges and universities. In many respects the SAT is a major source of information concerning the ability and preparation of college students, the quality of secondary education, and the extension of educational opportunity to minorities in a pluralistic society. Other admission tests, such as the ACT, also contribute significant and valuable information, but the SAT is more frequently discussed in matters of national or public policy. The SAT thus is much closer to being, in effect, a national or public examination for students leaving secondary schools and entering institutions of higher education.

During the 1980's, criticisms of standardized tests were frequently involved in the assessment of educational outcomes. Assessment was widely advocated for purposes of accountability and accreditation, and different forms and methods of assessing (as opposed to testing) student performance were part of the public demands placed upon schools and colleges. Criticisms of the SAT were particularly confused. Declining SAT scores were interpreted as evidence of the SAT's irrelevance to educational decisions, and variations in predictive validity coefficients were interpreted as proof of the SAT's technological obsolescence. Changes in the SAT's content were recommended by the SAT's more friendly critics, while complete abolishment was advocated by those who saw no merits whatsoever in either national testing agencies or their standardized products.

This chapter reviews the continuing national concern with access-and-equity issues in U.S. higher education and with public demands for accountability that stem from the declining quality of public education. The continuing interplay between a national quest for social justice, as reflected in educational opportunities and outcomes, and the public's interest in assessment (as a means of accountability), is a fascinating example of conflict in national character and temperament. Whatever the American national character, it has often been depicted as 'quite willing to enjoy the best of all possible worlds'.

The public interest in access and equity

American notions of equity stem from a long-standing concern with proportionate sharing. Regions, states, counties, and cities in the U.S. have often based their public expectations on concepts of 'democratic arithmetic'. If 2.6 percent of the nation's population resides within the State of Georgia, then Georgians are entitled to 2.6 percent of the nation's many benefits, advantages, and amenities. Such reasoning was quite pronounced in regional studies of the southern states during the 1930's and in the civil rights movement of the 1960's. Constituting 11.7 percent of the nation's population, black citizens understandably believe that they are entitled to 11.7 percent of the rights and privileges in which societal distribution is influenced by public policy.

As other minority groups became aware of their proportionate share of societal benefits and advantages, more frequent references were made to distributive justice – and institutions of higher education were increasingly perceived as the means by which such national objectives were attained. In much the same manner that national security and prestige had dictated federal aid to education in the National Defense Education Act of 1958, a national concern with equal opportunities shaped federal legislation in the 1960's. The Elementary and Secondary Education Act of 1964, the Higher Education Act of 1965, and the Civil Rights Act of 1964 were instruments of public policy that extended societal privileges, benefits, and advantages to previously excluded minorities.

For institutions of higher education, the societal extension of educational benefits and advantages can be seen as beginning in WWII with the enactment of the Servicemen's Readjustment Act of 1944. Further extension is seen in the work of President Harry S. Truman's Commission on Higher Education (1948). Explicit in that commission's report is the belief that public education should be accessible and should carry the individual as far as his or her 'native capacities' would permit. At least one half (49%) of the American people were capable of benefiting from fourteen years of formal schooling and at least one-third (32%) were capable of completing programmes of advanced liberal or professional education. The principal means by which such opportunities could be extended was two-year community or junior colleges that would provide a good mixture of general education with technical education. Precedents for such statements of faith in education can be seen in the Land-Grant College Act of 1862, in the development of American state universities (especially midwestern universities), and in the progressive movement of the early 1900's.

In the 1950's, questions of access (to higher education) became questions of national purpose and policy in many nations (Bowles, F., 1963). Secondary schools were part of an extended 'selection process' for institutions of higher education that were essential to national needs and objectives. The admissions process of selective colleges and universities became an intricate series of choices and decisions for individuals and institutions. In a publication recently re-issued by the College Board, B. Alden Thresher underscores the public's interest in a more effective system of college admissions. "Given the present assortment of 2,000 colleges, universities, and junior colleges, "he wrote, "what is the optimum way of

distributing students among them, taking the long-run public interest as the criterion?" (Thresher, B. A., 1966/1968, p. 83). Thresher was aware of the social and economic forces affecting public demands and expectations in higher education. As a result, he was dubious that either selective admissions or competitive recruiting would serve societal purposes well because students, institutions, and society have different objectives and expectations.

In the late 1950's and the early 1960's, as college enrollments expanded, selective admissions were momentarily regarded as a solution to limited facilities, faculties, and finances. The College Board took the lead in addressing the policy issues involved in limiting enrollments to accommodate institutional resources. Annual symposia, research conferences, professional journals, and the popular press demonstrated a remarkable consensus in their discussions of 'an impending tidal wave' of students and the nation's needs for campus facilities and classroom instructors.

The public interest in excellence

The individualistic features of the American national character have always included the expectation that individuals should excel. The discovery and development of talent was a particularly appealing aspect of higher education in its post-WWII years, and as school and college enrollments benefited from a rapidly increasing birthrate, the sifting and sorting of potential talent was a societal responsibility that colleges and universities managed well until overwhelmed in the late 1960's. Jeffersonian notions of a 'natural aristocracy of virtue and talent' were quite compatible with the nation's faith in education, and the testing of student aptitudes, abilities, and interests was in keeping with the public interest. John Gardner's (1961) book on *Excellence* was read as both a national and a personal challenge. Individual differences in athletics, music, art, and drama were widely recognized as natural talents that could be identified early and developed further. Creativity and ingenuity in science and other intellectual disciplines were increasingly recognized as special talents that should be discovered and encouraged. An emerging concept of developed abilities (such as those measured by the SAT) promised to set aside the irreconcilable differences of hereditarian and environmentalistic doctrines, and serious efforts were made to measure educational progress as a rational and developmental sequence in intellectual and/or academic competence. The optimism of that day was short-lived.

Social justice as access and equity

In the late 1960's, national thought and discussion were suddenly (and tragically) concerned with social justice. On college campuses student protests and faculty dissent produced an educational climate in which neither administrators, faculty, nor students could discuss access, equity,

and excellence as complementary concepts without resorting to contradiction. In their quest for excellence, colleges and universities were regarded by social critics as serving the screening and credentialing needs of society, business corporations, and government agencies. Excellence became a code word for elitism and for the exclusion of minority, women, and other non-traditional but deserving students. In their quest for equality of educational opportunity, schools and colleges witnessed a drastic shift from educational inputs to educational outcomes. Unwisely writing into the Civil Rights Act of 1964 a survey of equal educational opportunity, the U.S. Congress created a demand for instant results that schools and colleges could not supply. In their hurry to meet a congressional mandate, the survey researchers were entrapped by their methodological preferences and concluded (for the most part) that further federal funding was a waste of public resources. Instead of surveying the "availability of equal educational opportunities for individuals by reason of race, colour, religion, or national origin", the researchers analyzed the determinants of educational achievement, as measured by tests of verbal ability. The major conclusion, as reported in the news media and as heard by too many receptive ears, was to the effect that home, family, and community were more important to student learning than school facilities, teachers, classmates, and textbooks. Re-analyses and re-interpretations eventually clarified the confusions of the massive data collected, but not before irrevocable damage had been done in the world of public opinion and public expectations.

In 1973 the Carnegie Commission on Higher Education identified "the enlargement of educational justice for the postsecondary age group" as one of five main purposes of higher education. Following the education of individual students (purpose number 1) and advancing human capabilities in society (purpose number 2), the Carnegie Commission's version of social justice was listed as the third most important function of American colleges and universities. As purpose number 3, educational justice was ranked higher than pure learning (number 4) and the evaluation of society for self-renewal (number 5). In their general thrust and overall effect, the recommendations of the Carnegie Commission was quite compatible with other public attitudes, beliefs, and values of the day.

In retrospect, the emergence of equity as a dominant feature in public notions of social justice can be viewed as a poorly articulated progression from (a) public expectations of 'equal educational opportunities' to (b) strident demands for 'equal educational outcomes'. From there public thought and discussion turned to (c) concepts of 'equity' in educational opportunities and outcomes and eventually to (d) more realistic concepts of equity as 'fairness' in educational access, placement, instruction, evaluation, progression, graduation, and societal benefits or advantages. Concepts of equity as fairness are more in keeping with current concepts of cultural pluralism and give better promise to the future solution of educational problems. Unfortunately, the roadside is still cluttered with the debris of sociological and econometric notions of distributive effects that replaced a constructive 'psychology of individual differences' with a militant 'sociology of groups' (See Jencks, C., et al., 1979, as one example). In the process many schools and colleges found that they could not serve as instruments of worthy public policies without becoming political playgrounds.

The public interest in assessment

An increasing interest in the assessment of educational outcomes was clearly evident in the decade of the 1980's. National commissions and public leaders were apparently convinced that institutions of higher education could not be accountable without better evidence of congruence between institutional purposes and institutional effectiveness. Given the nation's previous reliance on testing, measurement, and evaluation, it is not surprising that the assessment of student outcomes should be widely advocated. Despite the inordinate concern with distributive effect in sociological and economic studies of occupational placement, starting salaries, lifetime earnings, socioeconomic status, and professional prestige, many studies were sensitive to the individual abilities of students and graduates as they climbed national career ladders. One of the more thorough studies (Sewell, W. H., Hauser, R. M., 1975) was explicit in its judgment of the importance of individual abilities: "However sophisticated our notions about social origins, it is not possible to give an accounting of the distribution of education, occupation, and income in the United States that excludes individual achievement and ability" (Sewell, W. H., Hauser, R. M., 1975, p. 184).

The 'assessment movement' of the 1980s is a function of the changing climate of public opinion in which institutions of higher education find themselves. A quarter-century of declining test scores, the necessity of teaching basic skills of literacy to thousands of college freshmen, and intense dissatisfaction with the general academic competencies of a large proportion of college students – plus embarrassing criticisms of the literacy of college graduates (enrolled in professional or graduate programmes) have convinced many public leaders and policy makers that outcomes assessment is necessary to ensure institutional and programme effectiveness and accountability to societal benefactors and sponsors. Assessment concepts and methods, as currently propagated in the U.S., have been influenced by the minimal competency testing movement of the 1970s, the futile debates concerning criterion-referenced and norms-referenced testing, and the nation's apparent love-hate relations with multiple-choice, mechanically scored, computer-processed tests such as the SAT. The U.S. news media – unable to accept the SAT as an empirically developed measure of verbal and mathematical ability and refusing to view standardized testing as a necessary technological innovation in an era of mass education – continues to give headlines to any and all studies that hint at Achilleian heels. Thus, they remain convinced that new and wondrous methods of assessing educational outcomes are available if only schools and colleges could be forced (by public opinion?) to use them. Interesting methods of assessment are in the making – but genuine, substantive assessment of general (and meaningful) educational outcomes is unlikely until national testing agencies supply 'an assessment market' with instruments and methods that will have the practicality, credibility, and fairness of traditional and/or empirically validated tests such as the SAT and ACT.

The assessment uses of the SAT are best demonstrated by the valuable information which the SAT (and similar tests) provides. No other source

produces as much relevant information about student abilities and achievement, and no other national organizations serve the public interest in the same way that the College Board, Educational Testing Service, and the American College Testing Programme do. Critics of the SAT seldom recognize that the intellectual development of individuals is still an essential purpose and function of higher education. Intellectual competence is still an expected outcome of a college education, and the SAT (with such tests as the GRE, LSAT, MCAT, etc.) is still the best measure we have of the intellectual/academic competencies needed in most fields of advanced, specialized, professional study. The verbal and mathematical abilities tapped by the SAT are still the best (general) indication of individual capabilities in the worlds of education and work. Also intellectual ability is still the best generalized measure of individual achievement. No one doubts individual differences in athletics, music, and art but many social critics deny individual (and group) differences in intellectual development.

In addition to information about the abilities and achievements of individuals, the SAT provides highly relevant information about the distribution of student abilities among institutions and programmes of higher education. Within statewide systems of public higher education, the distribution of students among institutions and programmes is particularly relevant to many questions of public policy. In the University System of Georgia, for example, the distribution of SAT scores among institutions lends credence to policy decisions that placed public institutions within commuting distance of the great majority (95%) of the State's population. Also implied by such distribution are the beneficial effects of individual choices and institutional decisions in the admission of high school graduates to the State's public colleges and universities (Fincher, C., 1986). The assessment uses and implications of the SAT may be summarized in the following manner: as a standardized test of academic ability, the SAT is, in many ways, a measure of basic academic competencies that are used extensively in undergraduate, graduate, and professional education. To no small extent, the test reflects the importance of vocabulary, reading comprehension, and mathematical reasoning in many different phases of conceptual learning. In addition to its usefulness as a general measure of intellectual/academic competence, the SAT has many advisement, counselling, sectioning, placement uses in academic decision making. Also relevant are the SAT's many uses in institutional research, planning, development, and evaluation (Fincher, C., 1986).

The SAT, perhaps more than any other nationally administered test, has contributed substantially to the clarification of public policies concerning access and equity. The policy issues concerning standardized tests, group differences, and assessment alternatives to traditional testing concepts involved, in many ways, data and information derived from the SAT or similar tests (Fincher, C., 1977, Fincher, C., 1979). For such reasons, the SAT, similar tests, or their modified versions should be used much more extensively in the assessment of educational outcomes – not less!

Access, equity, and assessment

The assessment of educational outcomes is a search for competence – not excellence! U.S. governors and other public officials demand assurances that high school graduates can read and write well enough to become responsible employees, voters, and taxpayers. Social justice requires that educational opportunities be accessible and that educational outcomes be equitable. The public interest requires better secondary preparation to ensure basic academic competencies that will permit high school graduates to take better advantage of postsecondary opportunities. Thus, it all suggests a remarkable convergence on one form or another of assessment. In the assessment of educational outcomes individual differences will continue to be the major source of variance in assessed outcomes. Neither institutional or programme characteristics, nor instructional facilities and methods will account for measured or assessed outcomes as well as the intellectual/academic development of individual students. The assessment of educational outcomes thus cannot be statistically significant, educationally meaningful, and publically creditable without consideration of individual differences in academic or learning ability.

Clearly it is in the public interest to assess individual differences in intellectual/academic competencies. It is also clear that social justice is best served in a society where individual talents, abilities, and achievements are assessed in a fair and creditable manner. Given systematic, objective, valid, reliable, creditable, and fair assessments of educational outcomes, individuals and institutions in higher education should be the beneficiaries. Assessed outcomes should serve the public interest in accountability and accreditation, as well as institutional interests in the improvement of undergraduate education. The assessment of individual competencies should serve social justice by contributing more substantially to self-understanding, academic and career planning, and personal management.

The assessment movement of the 1980's, as currently directed and supported, is without evidence that it can serve the larger purposes of access and equity in higher education. Social and economic forces (with their misplaced emphasis on distributive effect) continue to dominate educational thought and discussion; too many critics are obsessed with the trivialities of group differences in assessed outcomes; and too many advocates of assessment would be satisfied with simplistic before-and-after differences if they could be displayed positively in the news media. Should public opinion leaders and policymakers acquiesce to such notions of access, equity, and assessment, the public interest will not be served.

In the 1960's B. Alden Thresher wrote about "the range of social forces that caused many youngsters of high ability to shun higher education" (Thresher, B. A., 1966/1989, p. 11). In the 1990's we should study more closely the social and cultural forces that are barriers to access and achievement in a society with 3,500 institutions of postsecondary and higher education. We need to know more about the basic academic competencies of the nation's twelve million college students and why so many of them have not benefited more from a quarter-century of 'equal educational opportunity'.

References

Bowles, F., 1960, *Admission to College: A Perspective for the 1960's*, Princeton, NJ, College Entrance Examination Board.

Bowles, F., 1963, *Access to Higher Education*, Volume I, New York, Columbia University Press.

Cremin, L. A., 1989, *Popular Education and Its Discontents*, New York, Harper & Row.

Fincher, C., 1977, *Standardized Tests, Group Differences, and Public Policy*, The College Board Review, 103.

Fincher, C., 1979, *Using Tests Constructively in an Era of Controversy*, in: The College Board Review, 113.

Fincher, C., 1985, *Assessment Uses of the SAT in the University System of Georgia*, IHE Newsletter, December 1985.

Fincher, C., 1988, *Assessing Educational Outcomes: Are We Doing Good, Can We Do Better?*, IHE Newsletter, February 1988.

Fincher, C., 1986, *The Predictive Contribution of the SAT in a Statewide System of Public Higher Education*, in: Measures in the College Admissions Process: A College Board Colloquium, New York, CEEB.

Fincher, C., 1986, *Uses of the SAT in the University System of Georgia*, in: College Board Report, 86-5. New York, CEEB.

Fincher, C., *Trends in the Predictive Validity of the Scholastic Aptitude Test*, Princeton, NJ, Educational Testing Service (in press).

Jencks, C., et al., 1979, *Who Gets Ahead? The Determinants of Economic Success in America*.

Sewell, W. H., Hauser, R. M., 1975, *Education, Occupation, and Earnings: Achievement in the Early Career*, Studies in Population, New York, Academic Press.

Thresher, B. A., 1966/1989, *College Admissions and the Public Interest*, New York, The College Board.

Valentine, J. A., *The College Board and the School Curriculum*, New York, The College Board, 1987.

8 Issues in Assessment of Non-Objective and Objective Examination Tasks[1]

J. Izard

Introduction

When assessing progress in a school subject, teachers combine information from a number of sources in order to obtain an overall result for the subject. When deciding whether a student has passed in the year as a whole, the teachers combine results from a number of subjects. Examining authorities may use such information, or combine the information with external examination results when awarding certificates. This chapter examines the assumptions underlying such combination of scores, and explores the consequences of combining scores on separate tasks to obtain a total score in a subject, or a course.

To interpret such scores, a number of assumptions must be made. The assessment tasks are assumed to be relevant, and sampled from the knowledge and skills required by the student in his or her study of the subject. Further, it is assumed that the sample of assessment tasks is representative of all such tasks and that the sample reflects the balance of such tasks.

Generally, teachers combine scores on the basis of simple addition of the marks awarded to each of the components. In the simplest case, questions in a multiple choice test are scored correct or incorrect. Correct answers are assigned a value of 1 and incorrect scores are assigned a value of 0. The score on the test is the arithmetic sum of the values of the correct answers. It is assumed that each question is equally important in determining a total score and that it does not matter which questions contribute to that score. For example, all combinations of questions marked correct which give a score of 27 are regarded as equivalent.

In some cases, teachers are advised to set sufficient questions in each important topic of their course to enable the number of questions used in a topic to reflect the importance of the topic. If topic A is twice as important as topic B then topic A would have twice the number of questions as topic B. When tests are open-ended rather than multiple-choice the marking scheme may be more complex. For example, when questions are regarded as more

[1] The author is grateful for the assistance given by three separate groups of people in providing the data for these analyses. Dr Chris Haines, City University of London, not only provided the data for the first example, but also sent documents and letters with supporting information. The staff at the country High School in Victoria had been participant in the trials of the English Common Assessments. Dr Gunther Trost, Director, Institute for Test Development and Talent Research, Bonn, Germany, and his staff provided copies of the spatial perception sub-test and translated into English the instructions on the corresponding answer sheet. Master copies of other TMS sub tests were provided so that the copies used in this study could be produced. Dr Patrick Griffin, Dr Peter Ling, Mr Cliff da Costa, Mr Ralph Guice, and Dr Anthony Owens of the Assessment Research Centre at the Phillip Institute of Technology in Melbourne arranged the administration of the sub-tests and the collection of associated data. ACER's Clerical Service section coded the data for analysis.

important or substantial, they are usually assigned more marks. Incomplete or partially correct responses may be given some marks so that the number of marks for a task represents in some way the teacher's estimate of the extent of progress towards a perfect (i.e. correct) answer.

Teachers combine marks in non-test situations as well. When assigning marks to class work or study projects the traditional approach is to prepare a marking scheme which shares the marks available for the total task among each sub-task. For example, students presenting a group report may be given one-half of the marks for their investigation of the topic, including their searching of the library and the gathering of relevant details, and the remaining half may be given for the actual presentation in front of their classmates. Such a marking scheme may be interpreted by the students as assigning equal importance to the collection of information and its presentation.

Problems in combining marks

Radford draws our attention to the necessity of considering different perspectives, as follows:

> "*Lady Chatterley's Lover* was put in the dock in London as a corrupter of morals, but a reviewer of the magazine *Field and Stream* reported on it favourably in quite remarkable terms. The review states, in part:
>
> This fictional account of the day-to-day life of an English gamekeeper is of considerable interest to outdoor-minded readers, as it contains passages on pheasant raising, on apprehending of poachers, ways to control vermin and other chores and duties of the professional gamekeeper. Unfortunately one is obliged to wade through many pages of extraneous material in order to discover and savour those sidelights on the management of a Midland shooting estate. This book cannot, however, take the place of J. R. Miller's *Practical Gamekeeping*."
> (Radford, W. C., 1969, p. 150)

There are several perspectives to consider when combining marks. One perspective denies meaning to a combined score unless there is internal consistency between the various tasks, at least with respect to intercorrelations between scores on those tasks. In a test development context, researchers use procedures such as reviewing discrimination indices and item-characteristic curves and factor analysis when deciding whether a question 'belongs' to the test or not. From this perspective questions judged inappropriate are deleted from the test. However, deleting questions on this basis can result in the final form of the test being less representative than specified originally.

In conventional multiple-choice tests individual questions correlate with the total score to varying extents. If the teacher conducts an item analysis to detect questions which are inconsistent with the rest of the test, then the questions that are retained may be considered to be internally

consistent. However excluding some questions may make the test less valid as a measure of the extent of coverage of the course being studied. For example, in mathematics tests some questions may involve quantitative skills, others may involve spatial skills, and others may require verbal skills to tease out the meaning of the problem before applying other skills. In a test with a substantial proportion of quantitative questions and a small proportion of spatial questions, the latter may not correlate with the rest of the test to the same extent as the quantitative questions. Consideration of the item statistics may lead to the exclusion of the spatial items with an apparent improvement in the internal consistency of the test. But the revised test now has no questions on a significant part of the mathematics course as originally designed. Improved internal consistency has been achieved at the cost of validity, even if the remaining questions can be used to develop a new test to provide a separate spatial score. However this does not solve the problem. The marks for the quantitative part still have to be combined with the spatial part to give a total score on the course as a whole, and the procedure for obtaining such a total score needs to be chosen carefully.

From another perspective, the meaning of combined scores is defined arbitrarily, regardless of the patterns of intercorrelations between scores on the tasks. Marks on Mathematics may be combined with marks on Science, English and History to decide on eligibility to enter a tertiary course. This approach is not limited to the combination of subject marks to decide an overall score. The same approach often exists within subjects. For example, marks on two Mathematics examinations may be combined to give an overall Mathematics score. In English for the Victorian Certificate of Education, several separate assessments are made. These marks are combined to obtain the mark for English. The meaning given to a composite score in such a situation may be obscure but the procedures are agreed either by legislation, regulation, or actual practice, and such procedures are used by many examination boards. One question of interest is, 'Does each assessment have the same value in deciding the total mark?'. Another question is, 'Does the assessment for each part have an actual weight in the final score equal to the intended weight in the final score?'.

When combining marks, teachers assume that simple addition of marks for components will give the nominal weightings as actual weightings. For example, when adding the scores on English and Mathematics it is expected that a total score will give equal weight to each subject. No account is taken of the variation within each set of scores, or of the extent of covariation. For example, teachers marking an English paper may use a limited number of categories in assigning marks, while those marking the Mathematics paper may assign marks in such a way that the entire scale from zero to the maximum is used. Combining such pairs of marks will give more emphasis in the total to those marks which have a wider spread.

Further, many teachers do not realise that adding marks on like subjects may influence the total scores in a different way to adding marks on subjects which differ. No account is taken of overlap between subjects in following the rules for aggregation. It may be that scores on two Mathematics subjects are added and regarded as equivalent to scores on English and Music, or some other pair of subjects.

In order to examine the contribution of the parts to a total score it is necessary to be aware of the spread of scores within each part, as well as the extent of the correlation between the parts. Scaling procedures are required for the separate assessments within a subject to obtain an aggregate for the subject as a whole, and for combining separate subjects when calculating an aggregate for selection purposes.

Illustrations with real data

Three illustrations are presented. In the first example, results on several subjectively-scored sub-tasks are aggregated to give a score on a project which contributes to a final result in a degree course. In the second example, a pencil-and-paper examination task is combined with results on four other tasks. The final example involves aggregation of three separate measures with low intercorrelations in order to establish a rank order that may be useful in the selection context.

Example 1

In a paper presented at the 4th International Conference on the Teaching of Mathematical Modelling and Applications, Haines (Haines, C. R., 1989) described how projects in the final year of the mathematical science programme were assessed. These projects had been carried out over the first two terms. Each student had undertaken an individual project working under supervision within specific agreed aims and objectives. The assessment was made by the supervisor according to a marking scheme and a short report was written. The report including the mark was subject to the scrutiny of a moderator and an agreed mark was then submitted for the project. There was no oral presentation by the student.

The marking scheme used from 1986 until 1988 allocated marks in the following proportions:

Summary or abstract	10
Problem statement/introduction	20
Content	50
References/data sources	10
Presentation	10
	100

(The marking scheme was modified on the basis of a personal communication of C. R Haines of August 10, 1989, the proportions becoming 10, 16, 50, 8, and 16 respectively.)

Haines explored various ways of combining the data. In addition to using the traditional sum of scores model he applied a product model (after Hall, G. G., 1983) which had the effect of giving a combined score of zero if any one of the categories was assessed as zero. His exploration suggested

that, in practice, very few categories are given a zero mark and as a consequence, the aggregate scores were similar whether the sum model or the product model was applied.

Using the data from Haines' study one can examine the contribution of each aspect or category to an aggregate, provided that one assumes that the marking scheme was applied in the same way for each candidate even though the content of each project varied. Table 1 shows the summary statistics for each aspect of the marking scheme.

Table 1 Summary statistics for categories (Haines 1986-88 data)

N=52	Mean	Standard Deviation
Summary or abstract	7.15	1.51
Problem statement/introduction	3.60	0.77
Content	18.14	4.05
References/data sources	3.42	1.04
Presentation	3.89	0.94

Table 2 shows the correlations between each of the category scores. Haines recognised that failure to use the complete scale in any category would affect the correlations obtained, but argued that, provided the assessment was criteria based within well defined boundaries, the assessments would be appropriate descriptions of the actual merit of the projects.

Table 2 Correlation coefficients for categories (Haines 1986-88 data)

N = 52	Problem	Content	References	Presentation
Summary	0.67	0.69	0.36	0.36
Problem		0.69	0.36	0.37
Content			0.44	0.53
References				0.65

One way of looking at the actual contribution of each part of the project is to use multiple regression (Darlington, R. B., 1968). Table 3 shows the beta weights for each of the parts when the aggregate is calculated by adding without scaling. The actual weights differ from the intended weights.

Table 3 Regression beta weights for categories with traditional aggregate (Haines 1986-88 data)

N = 52 (traditional aggregate)	Beta weight (Rsq = 0.97)
Summary or abstract	0.22
Problem statement/introduction	0.11
Content	0.59
References/data sources	0.15
Presentation	0.14

Another way of looking at the same information about the parts of each project is to aggregate using the Partial Credit Model (Wright, B. D., Masters, G. N., 1982) and analyse contributions to this aggregate in the same way as before. Table 4 shows the beta weights for each of the parts for such an aggregate.

Table 4 Regression beta weights for categories with Partial Credit aggregate (Haines 1986-88 data)

N = 52 (Partial Credit aggregate)	Beta weight (Rsq = 0.95)
Summary or abstract	0.26
Problem statement/introduction	0.22
Content	0.25
References/data sources	0.29
Presentation	0.25

The aggregate obtained using the Partial Credit Model is more balanced in the sense that each of the parts in the assessment make a similar contribution. However, the weights still differ from the intended weights.

Example 2

In the Australian State of Victoria a different approach to assessment at Year 12 level is being introduced (Masters, G. N., 1989). The proposed use of a number of common assessment tasks (CATs) in each field of study (subject) has resulted in considerable public debate, and a report on the proposal has just been published (McGaw, B., Eyers, V., Montgomery, J., Nicholls, B., Poole, M., 1990). Within a school subject four tasks will be given. In English these may be a writing folio, an investigative report, oral communications, and text response. Some may be assessed within the school using a regional moderation process to seek comparability of assessments across different schools. The aggregation of these assessments (after scaling) by the Victorian Curriculum and Assessment Board (VCAB) for each individual provides the overall assessment in that subject.

Within a single school, teachers need to be aware of the effects on the aggregate score of the assessment strategies used for each common assessment task within a particular subject. The data for this example came from the 1989 trials of the CATs in English in one of the 11 trial schools for the new approach. (It should be noted that the data are for those students who completed all five tasks, whether they were candidates for the traditional Higher School Certificate (HSC) or the alternative Schools Year 12 and Tertiary Entrance Certificate (STC). Three of these were excluded from the analysis because one had the maximum score and two had the minimum score. Other students excluded were those with incomplete records (16), those who withheld their permission for release of results to their school (3), and those who were given special consideration under provisions for disadvantage(2).) There were five common assessment tasks

in English given to Year 12 in 1989.[2] These were:

- *Investigative Report (IR)*, a presentation of a personal point of view on a topical issue, drawing on at least three media texts and one other source such as a literary text, a reference text, or an interview or survey (July submission);
- *Writing Task (WT)*, a piece of writing on a specified theme, in response to stimulus material and completed in two hours under examination conditions at the end of Term 3 (September);
- *Text Response (TR)*, a selection of the student's two best pieces of writing on four texts studied during the course (October submission);
- *Writing Folio (WF)*, to contain five pieces of finished writing with at least three in different genres together with evidence of drafting with two of the pieces (October submission);
- *Oral Communication (OC)*, at least three oral presentations of different types, some related to work requirements in the course, such as a communication project, another related to other common assessment tasks such as the investigative report or the text response (November submission).

Results on each of the common assessment tasks had been rescaled to a five-point scale by VCAB. Table 5 shows the summary statistics for scores on the five separate assessment tasks for the 67 students retained in the analysis.

Table 5 Summary statistics for common assessment tasks (1989 Year 12 data)

N=67	Mean	Standard Deviation
Investigative Report (IR)	2.99	1.29
Writing Task (WT)	2.57	1.02
Text Response (TR)	2.76	1.10
Writing Folio (WF)	3.02	0.95
Oral Communication (OC)	3.08	1.13

Table 6 shows the correlation coefficients obtained between scores on the five separate assessment tasks for the 67 students retained in the analysis. Table 7 shows the corresponding coefficients for all 11 trial schools (from McGaw, B., et. al., 1990, p. 25). Tables 6 and 7 show similar patterns of correlations although it must be noted that some agreement would be expected because the single school was one of the 11 schools.

[2] For 1990 the number of tasks has been reduced to four. The Writing Task has been deleted, with the Text Response becoming the task undertaken as a written examination. Those from the nothern hemisphere should note that academic year for Victorian schools commences in February (after the summer holidays and finishes in December.

Table 6 Correlation coefficients for English assessment tasks (1989 Year 12 data, one school)

N = 67	WT	TR	WF	OC*
Investigative Report (IR)	0.38	0.59	0.56	0.39
Writing Task (WT)		0.41	0.43	0.37
Text Response (TR)			0.80	0.61
Writing Folio (WF)				0.49

*Oral Communication

Table 7 Correlation coefficients for English assessment tasks (1989 Year 12 data, 11 schools)

	WT	TR	WF	OC*
Investigative Report (IR)	0.43	0.63	0.68	0.54
Writing Task (WT)		0.47	0.53	0.46
Text Response (TR)			0.80	0.58
Writing Folio (WF)				0.60

*Oral Communication

When contributions of each CAT to the English score awarded by VCAB are assessed for data from the one school using multiple regression, the beta weights shown in table 8 are obtained.

Table 8 Regression beta weights for English assessment tasks with traditional aggregate (1989 Year 12 data, one school)

N = 67 (traditional aggregate)	Beta weight (Rsq = 0.98)
Investigative Report	0.26
Writing Task	0.22
Text Response	0.31
Writing Folio	0.20
Oral Communication	0.27

If the Partial Credit Model is applied to the same data for the CATS another aggregate is obtained, with beta weights as shown in table 9. The similarity of beta weights is partly a function of the scaling of each CAT by VCAB before issuing scores to the school, and partly because the aggregate English score awarded by VCAB was obtained by a similar procedure. However the weights for the aggregate obtained using the Partial Credit Model are more similar than the corresponding weights for the other aggregate.

Table 9 Regression beta weights for English assessment tasks with Partial Credit Model aggregate (1989 Year 12 data, one school)

N = 67 (Partial Credit aggregate)	Beta weight (Rsq = 0.99)
Investigative Report	0.29
Writing Task	0.21
Text Response	0.27
Writing Folio	0.24
Oral Communication	0.26

Example 3

In a paper presented at the 15th International Conference of the IAEA, Izard and Griffin (Izard, J. F., Griffin, P., 1989) described a pilot study using sub-tests from a test used in a nationwide testing program for admission to Medical Schools in West Germany (Trost, G., 1988). The three sub-tests from this *Test for Medical Studies* (TMS) were memory for figures, differentiated visual perception, and spatial perception. In the first sub-test, candidates were required to memorize information and to distinguish this information from other similar information when presented some time later. The second sub-test assessed differentiated visual perception with pattern recognition tasks, and the spatial perception sub-test presented tasks involving different perspectives of cables. Means and standard deviations for each sub-test are shown in table 10 (1989 data) and table 11 (1990 data).

Table 10 Summary statistics for sub-tests (1989 data)

N = 177	Mean	Standard Deviation
Pattern Recognition	10.72	3.21
Perspectives of Cables	12.08	3.61
Memorizing Figures	8.40	3.51

Table 11 Summary statistics for sub-tests (1990 data)

N=401	Mean	Standard Deviation
Pattern Recognition	9.51	3.32
Perspectives of Cables	10.62	3.80
Memorizing Figures	7.56	3.40

While these sub-tests appeared to have a common spatial element, the scores on the sub-tests did not have high between-sub-test correlations. The correlations between the sub-tests are shown in table 12 (from Izard, J. F., Griffin, P., 1989) and table 13 (Data collected in 1990).

Table 12 Correlation coefficients for sub-tests (1989 data)

N = 177	Perspectives of Cables	Memorizing Figures
Pattern Recognition	0.30	0.02
Perspectives of Cables		0.18

Table 13 Correlation coefficients for sub-tests (1990 data)

N=401	Perspectives of Cables	Memorizing Figures
Pattern Recognition	0.36	0.23
Perspectives of Cables		0.30

Again, the simple aggregate of unscaled scores gives one of the tests more weight in the total as shown in table 14. The Partial Credit Model aggregate gives both sub-tests comparable weights as shown in table 15.

Table 14 Regression beta weights for sub-tests with traditional aggregate (1989 data)

N = 177 (traditional aggregate)	Beta weight (Rsq = 0.64)
Perspectives of Cables	0.63
Memorizing Figures	0.50

Table 15 Regression beta weights for sub-tests with partial credit aggregate (1990 data)

N = 401 (partial credit aggregate)	Beta weight (Rsq = 0.83)
Perspectives of Cables	0.56
Memorizing Figures	0.56

Implications

Aggregation of scores, whether from components of a single subject, or from combinations of subjects, makes an assumption that the combined score has meaning. Aggregating unscaled scores distorts the results in various ways and cannot be defended. In each of the examples given above, regression weights for components were closer together when a scaling approach was used (in this case the Partial Credit Model). However to assume that marks for the various components have the weights implied by the teachers' marking schemes is unwise. Use of the Partial Credit Model did not adjust the weights towards the teachers' marking scheme, but did provide approximately equal weightings for the components.

When scores on tests are combined in some way to establish a rank order of candidates for selection purposes, it is difficult to give meaning to the resulting composite score. When the selection criteria require that results on several measures are combined, it is important to ensure that an appropriate scaling approach is used to avoid distorting the rank order in some unspecified (and probably unknown) way.

The investigation of these examples focused on weights within particular assessment contexts. It is likely that similar problems occur when examining boards make comparisons between candidates in separate cohorts.

References

Darlington, R.B., 1968, *Multiple Regression in Psychological Research and Practice*, in: Psychological Bulletin, 69, pp. 161-182.

Haines, C.R., 1989, *Project Assessment for Mathematicians*, Paper Presented at the Fourth International Conference on the Teaching of Mathematical Modelling and Applications, Roskilde.

Hall, G.G., 1983, *The Assessment of Modelling Projects*, in: I.S. Berry, et al. (Eds.), Teaching and Applying Mathematical Modelling, Chichester Ellis Horwood, pp. 143-148.

Izard, J.F., Griffin, P., 1989, *Spatial Tests for Tertiary Selection: A Pilot Study*, Paper Presented at the Fifteenth International Conference, International Association for Educational Assessment, Sydney.

Masters, G.N., 1989, *Selection for Higher Education and the Reform of the Senior Secondary School*, in: Search of a Working Compromise, Paper Presented at the Fifteenth International Conference, International Association for Educational Assessment, Sydney.

McGaw, B., Eyers, V., Montgomery, J., Nicholls, B., Poole, M., 1990, *Assessment in the Victorian Certificate of Education: Report of a Review Commissioned by the Victorian Minister for Education and the Victorian Curriculum and Assessment Board*, Melbourne, Curriculum and Assessment Board.

Radford, W.C., 1969, *Some Changing Attitudes to and Views on Evaluation and Measurement*, in: The Australian College of Education, Educational Measurement and Assessment, Melbourne, The Australian College of Education, pp. 150-163.

Trost, G., 1988, *A Nationwide Testing Program for Admission to Medical Schools in West Germany*, Paper Presented at the 14th International Conference of the International Association for Educational Assessment, Iowa City (Iowa).

Wright, B.D., Masters, G.N., 1982, *Rating Scale Analysis*, Chicago, MESA Press.

9 Determining Equivalence among European Educational Qualifications

S. Johnson

Introduction

In the late 1980's the European Commission launched a range of Europe-wide cooperative initiatives in the field of education and training, recognising the crucial role that education and training will play in the achievement of a successful European internal market. As identified by the Commission itself (Commission of the European Communities, 1989a), the principal programmes are the following: ERASMUS, intended to promote student mobility; COMETT, planned to further cooperation between universities and enterprises in the field of new technologies; LINGUA, to promote foreign language teaching; EURYDICE and ARION, to strengthen cooperation and information in the field of education; PETRA, to improve the preparation of young people for adult and working life; EUROTECNET, to develop vocational training in the field of new technologies; CEDEFOP, to promote the development of vocational training through information, research and cooperation; IRIS, to develop equal opportunities and vocational training for women; and YOUTH EXCHANGES, a young workers exchange programme. Taken together, these initiatives address all levels of education, from primary to tertiary, both vocational and academic.

It is expected that such sponsored cooperative ventures will produce the most effective exchange of knowledge, skills and ideas at the European level. The initiatives, particularly those involving student, teacher, and worker mobility should, moreover, simultaneously encourage increased cultural awareness and mutual tolerance amongst different European nationals. Individuals should benefit from their participation in terms of personal development and future marketability; the EC hopes to benefit in terms of future economic health and social cohesion.

It is in the context of the increased mobility of students and workers that the equivalence issue arises, and two kinds of certificate recognition have been distinguished: professional recognition and academic recognition. The former describes the situation in which summative qualifications are being judged as appropriate or otherwise for the exercising of a profession or trade. Should a qualified pharmacist in one member state wish to work as a pharmacist in another then a judgement would be needed about the equivalence of pharmacy qualifications in the two countries concerned. Academic recognition, on the other hand, refers to the situation where students with educational experience gained in one member state wish to continue their education in another. Here, the equivalence issue arises in the context of suitability for admission to, and appropriate placement within, the educational system of the other country. It is the diversity of educational systems and their certificates within Europe which renders the question of equivalence so problematic. Academic

recognition is a central issue for the Commission's ERASMUS programme, whose main focus is cross-frontier student mobility. Students in an academic institution in one member state can opt for study periods of at least three months duration in a host institution in another member state, during the normal course of their study towards a particular level of qualification. Such mobility is encouraged through financial assistance, but also, and equally importantly, with the possibility of formal recognition of the study abroad by the home institution.

It has always been possible for students to move from one institution to another in the middle of their undergraduate or graduate studies within their own countries or across frontiers. But such individuals would often do this at great personal cost, particularly in those cases where previous study could not be credited in any way. The possibility under ERASMUS of academic recognition for periods of study in other member states is therefore a welcome innovation, already proving attractive to thousands of students. In 1987/88, the first year of operation, 4000 students took advantage of the study-abroad option; by 1989/90 the figure had already risen seven-fold to 28000, 2% of all higher education students in the EC (Commission of the European Communities, 1989b). By 1991/92 the plan is for the proportion of study-abroad students to rise to 10%.

Perhaps the most exciting, and certainly the most challenging, feature of the ERASMUS programme is the European Community Course Credit Transfer Scheme (ECTS), which has just completed its first pilot year of operation (Commission of the European Communities, 1990). Under this scheme, students will, in principle, qualify for academic recognition of their study periods abroad in the form of formal credit transfer. In its five pilot subject areas, Business Administration, Chemistry, History, Mechanical Engineering and Medicine, ECTS has chosen to adopt the American education system as its general model, with modifications. Courses are subdivided into identifiable units, and units are allocated a number of credits to reflect their length, level and status. Students will be expected to follow a programme of work leading to a 60-credit gain in any full year of study. At the present time the scheme is inevitably based on mainly bilateral agreements between individual institutions, but if successful its scope will broaden.

Academic recognition is not a new concept, of course. Most schools and colleges around the world have from time to time had to make judgments about the academic credentials of foreign applicants. But it is arguably within the American higher education system that the greatest amount of experience in this field has been accumulated, not only through credit transfer between stateside institutions but also in relation to the admission and 'advancement placement' of non-US nationals. Hence the decision to base ECTS on a modified US model. In addition, there are more than 750 'international schools' around the world, whose pupils and teachers are of many different nationalities (ECIS, 1990). They also have had to think more than most about this matter.

It is with the concept of academic recognition that this chapter is concerned. Since academic recognition ultimately rests on assumptions and shared agreements about equivalence in terms of curriculum exposure and standards of achievement, the chapter in particular focuses on these issues,

highlighting implications for any programme of equivalence determination within Europe. For illustration purposes, reference is made to the educational systems of three particular countries. One is an EC country, the UK, in whose system the author was educated, the second is a non-EC European country, Switzerland, whose system the author is currently studying at close quarters, and the third is a non-European country, the US, in whose higher education system the author currently works, and on whose credit-transfer system ECTS is being modelled. The educational systems of these three countries alone, and in particular their evaluation policies and practices at national level, vary sufficiently to indicate the kind of diversity to be found among the countries of Europe as a whole.

Academic recognition

When college applicants are agreed to have equivalent educational experience, whether certificated or not, we generally mean by this that they can be considered similarly well-prepared for further study of the kind proposed. Similarly well-prepared that is in terms of curriculum exposure and standards of achievement, excluding considerations of personality, motivation, non-native language competence, and so on. The nature of the further study will determine equivalence to the extent that different educational experiences might be considered interchangeable in some circumstances but not in others.

In the American system, for instance, the basic requirement for college entry for American students is a high school diploma, which the majority of the age-population will hold. Selective institutions usually, in addition, demand some pre-determined minimum score on the mathematics/English Scholastic Achievement Test (SAT) or similar. Because there is no standardisation of curriculum at high school level in the US, either by federal or state dictate or through external examinations, colleges cannot impose subject-based conditions on entry. In principle, any high school diploma is as good as any other, irrespective of the institution which awarded it and irrespective of the profile of courses detailed in the transcript. The subject mixes are treated as interchangeable. But when we come to credit transfer between colleges and universities the picture sometimes changes. A course studied in one college might qualify for general credit transfer to another, but it might not be considered suitable to allow a specific course exemption. That would depend on the similarity, or degree of equivalence, between the two courses.

In Switzerland the qualification which guarantees entry into higher education is the Federal Maturitatszeugnis, otherwise known as the Certificat de Maturité or Attestato di Maturità. This certificate is gained through national examinations on a Confederation determined syllabus, after 12-13 years of primary and secondary education. It is held by only around 15% of the age-population. Pupils take a broad range of core subjects, but can in addition specialise in one of five areas: Latin and Greek, Latin and modern languages, Mathematics and sciences, Modern languages, and Economic sciences. Whatever the specialism the certificate

allows unrestricted entry to any faculty in any of Switzerland's universities or polytechnics, although in practice the degree of specialism in the Maturità will lead to a logical follow-on education at the higher level.

In the UK the situation is different again, with a greater degree of subject specialisation beginning much earlier in the system. Here, the usual ticket for entry to higher education for British students is the General Certificate of Education at Advanced level. This is a subject-specific certificate, unlike those of Switzerland and the US, achieved through external examination after 12-13 years of primary and secondary education. Currently, around 20% of the age-population will hold at least one A-level certificate; just over 10% will have passes in three subjects, the normal requirement for the more selective institutions. Intending students apply for admission to specific departments in their chosen universities and polytechnics, and these departments often impose both general and subject-specific entry conditions. Highly selective departments generally demand high-grade passes in three subjects at A-level; others would require less in one or both respects. In most cases the subject combination would need to be relevant for the intended field of study, and in all cases applicants must hold passes in mathematics and english at a lower level.

These examples of operational selection criteria suggest that there are some general qualities which education is considered by and large to foster, over and above the transmission of subject-specific knowledge and the development of subject-specific abilities and skills. We might cite intellectual curiosity and tolerance, analytical abilities, study skills, social skills, and so on. On the other hand, perhaps the mere fact of having stayed within the educational system for so many years simply confirms a certain level of intellectual power, and demonstrates an ability to persevere to a successful conclusion with whatever the task in hand.

One could argue at length about the general nature of the educational process, and about its worth to employers and others. The point of relevance here is that summative qualifications at a particular level in the educational system are frequently considered interchangeable as confirmations of past achievement and/or as indicators of future intellectual performance, regardless of their nature or degree of specialisation. Despite often gross differences in terms of subject matter, equivalence is assumed. In other circumstances these same qualifications would not be considered equivalent, the subject matter involved assuming greater, even paramount, importance. In most circumstances equivalence considerations would extend to include standards of achievement.

At its very simplest, equivalence can be defined, rightly or wrongly, in terms of duration of full-time education. In other contexts equivalence would mean more than this, and would extend to cover the nature and amount of subject-specific education. This might further be defined in curriculum terms, such as similarity of content coverage, maybe even taking account of the profile of abilities and skills addressed during the education leading to the certificate. Finally, equivalence can be more narrowly focused still, involving judgments about comparative standards of achievement. The more narrowly we attempt to define equivalence the more complex the issues appear, particularly when we extend consideration beyond the confines of a single national system.

American experience with academic recognition is illuminating as we contemplate the problem within Europe. Important features of the American education system are its generally open access, its credit-based nature and curriculum flexibility, and, most important of all in this context, its credit transfer arrangement. Most American colleges and universities accept students on temporary or permanent transfer from other institutions, at any point in their undergraduate or graduate careers. In addition, many colleges have 'study abroad' programmes, allowing their students to spend a semester or year studying in another country. In principle, though not necessarily in practice, credits earned in any accredited institution, stateside or abroad, transfer with the student. Thus, for the most part students have been able to move from one institution to another relatively freely, without personal penalty. The system has continually extended to accommodate the growing number of students from Europe and elsewhere wishing to study for American qualifications, most of whom qualify for 'advanced standing'.

More than twenty years ago, to give guidance to college registrars faced with applicants holding foreign pre-university certificates, the American Association of Collegiate Registrars and Admissions Officers began publication of its *World Education Series*, comprising descriptions of the educational systems of different countries around the world with guidelines on appropriate placement of foreign students within the US system. Most recently, this same body has published an *International Academic Credentials Handbook* (AACRAO, 1988, 1989), which draws on information from the World Education Series. The handbook looks in detail at the variety of certificates which are awarded at the different levels of education within each country, and includes exemplars of the certificates themselves, with helpful explanations and occasional language translation of critical phrases. Other, less comprehensive and less rigorous guides include *The Country Index* (Sepmeyer, I. H., 1986), and the *Handbook on the Placement of Foreign Graduate Students* (NAFSA, 1984).

For the most part, these publications limit themselves to describing the educational systems in each country, not always an easy task in itself given the complexity involved in many cases, and to drawing very general guidelines on equivalence in terms of the age at which certificates are normally awarded, and the duration and broad nature of the education involved. Placement recommendations are offered, which suggest appropriate amounts of credit to award for various qualifications.

Inevitably in the circumstances of international diversity, admissions officers and registrars are still left to use their own initiative to a very great extent. Thus when applicants request credit transfer, and especially when they appeal for exemption from specific required courses on the basis of their previous education in another institution at home or abroad, teachers would usually be consulted for decisions. Often decisions are made on the basis of course title alone. 'College Algebra', 'Calculus', 'History of the US', 'Introduction to business', and indeed most foundation-level courses with familiar titles, would be fairly readily accepted on this basis. Higher-level courses, especially those with relatively unfamiliar titles such as 'Research Methods', 'Controversial economics' or 'Cross-cultural awareness', would be considered with greater caution, the degree of caution depending on the reputation of the institutions involved.

Where the relevant document is available, judgments about appropriateness for credit transfer and/or for course exemption might involve a comparison of subject coverage at the level of the mandatory, though generally brief, course description supplied by the other institution. The course description would usually include a statement of teaching objectives, a timetable of work, text book title and evaluation criteria (weightings given to the various evaluation components). Copies of course assignments or examination papers would not be available for perusal, even had the individual concerned the time and inclination to request, await and make use of these.

As far as standards of achievement are concerned, it seems that, despite the extremely naive inferences about grade correspondence between the US and other national systems offered in some of the guideline publications (NAFSA, 1984, Sepmeyer, I. H., 1986), few institutions actually transfer grades. However, many do take grades into account when deciding whether to accept credit transfer or not, with a wide variation in practice (Nisbet, C., 1980), and in doing so are making assumptions about relative standards of achievement in the absence of any supporting evidence. But how can decision-makers be helped further in their task? How can guidelines be improved so that more sophisticated information and advice can be offered? How can the current level of injustice being done to individual students be minimised? These are the questions we must address if Europe's credit transfer system is to operate more efficiently and more fairly. Let us adopt the best of American practice, but let us also recognise the flaws in the system and look for solutions. This is the challenge we are offered at this point in time.

Establishing curriculum equivalence

Whether, and to what degree, curriculum equivalence is important can depend on the nature of the subject concerned, and on the particular circumstances under which equivalence is being determined. In some circumstances, for example, all introductory computer courses might be considered interchangeable with each other, in the sense of having introduced the student to some basic concepts determining 'computer literacy'. Even major differences such as one course focusing on computer programming and another featuring rather the use of applications packages would be considered unimportant. In other circumstances such differences would be considered too great for equivalence to be accepted. A student holding credit for a programming course would be considered less suitably prepared for an advanced course on management information systems than a student with credit for an introductory course focusing on the use of applications packages. Here we have an example of a lack of 'functional equivalence'. Different kinds of course might be considered educationally equivalent in some respects, but in the circumstances of preparation for a particular kind of further education they would not be equivalent enough.

Similar examples can be given in other subjects, such as history. Suppose, for instance, that one student had successfully completed a course in European history whereas another had followed a course of the same

duration and general level of education in, say, American history. Despite the difference in factual content, would the respective educational experiences of these two students in history be considered equivalent in any sense? History teaching and learning are not concerned uniquely with knowledge transfer; historians, like any other professionals, call on qualities other than, and in addition to, their ability to memorise and to recall facts – skills of information analysis and communication for example. If a 'feel' for what is involved in studying history is the focus, then any history course would be considered interchangeable with any other; all history courses would be equivalent. Only if the courses were being judged for functional equivalence, in the sense of adequacy of preparation for a specific subsequent course, might the content difference matter.

Courses with identical content bases can similarly be considered equivalent or non-equivalent curriculum terms, to the extent that they emphasise to differing degrees different aspects of subject ability. To give an extreme example, of two courses with the same content syllabus one might be concerned exclusively with the acquisition of knowledge whereas the other might rather focus on the development of information skills, the content base merely serving as a vehicle for skills development. The courses would use the same factual material, but they would be different courses nonetheless. Would their educational outcomes be considered interchangeable in any circumstances? If so, in which circumstances?

Such questions are not easy to answer. If they can at least be asked then that is progress. For it is sometimes the case that differences in emphasis of this kind are not readily discernible. Course objectives, loosely worded as they often are, and even week-by-week timetables of work are not guaranteed to reveal these features. Where differential emphases reflect the personal interests and teaching style of individual teachers then they are unlikely to be readily apparent, and in any case little account can be taken of them in practice. Individual students will simply be more or less advantaged or disadvantaged when they come to pursue other courses given by other teachers. On the other hand, where differences in emphasis are institutional, or even national, then these can at least be investigated and recorded. Such investigation need not involve observations of teaching practice, but rather would concern itself with a scrutiny of teaching materials, assessment materials and methods, and evaluation practices and policies. It is evaluation policies that determine what is actually credited during a course, as opposed to what is claimed to have been taught. Even where course descriptions suggest close similarity in content coverage and skill/ability emphasis, the structure and forms of associated tests and assignments, along with mark allocations, and aggregation policies, can suggest the opposite. Evaluation validity is the key term here.

Within the UK higher education system, such analyses of examination structures have indeed revealed differences in skill/ability emphasis in the same subject degree courses of different universities (Johnson, S., 1988). Formal mark schemes are not generally available at this level of education, and so could not be scrutinised. However, reviews of mark schemes used in external examining at the level of secondary education in this country confirmed that these provide further and final evidence of these kinds of difference (Johnson, S., Cohen, L., 1983, Johnson, S., 1988).

There are clear implications for the situation in which specific individual courses are being evaluated for relative suitability as preparation for further work. Ideally, judgments about curriculum equivalence should be based on a comparative review at a level deeper than that of course titles, or even of stated objectives and syllabus. In practice this rarely happens. As noted earlier, brief course descriptions would often be the best basis available on which a teacher in a US college might make such decisions. Thus it is that judgments are made about whether 'Introduction to finance' from Institution A, 'Beginning Finance' from Institution B, or 'Finance and Accounting' from Institution C, would substitute appropriately for 'Finance', a required course for a business major in Institution D. Without more detailed course information the teacher makes as valid a decision as possible.

Whatever the degree of validity of individual exemption decisions, students with course exemptions on the basis of credits for 'equivalent' rather than identical courses will begin their higher-level courses with different kinds and amounts of subject experience. This situation of inequity in preparation inevitably leads to one of inequity in ability to make effective use of the new learning opportunity. It can also lead to inefficiency in the system, slowing down the pace of education.

Establishing equivalence with respect to standards of achievement

Courses can differ in terms of their duration, their content base, the range of abilities and skills fostered, and the relative emphasis given to each of these abilities and skills. They can also differ in terms of the level of demand imposed, and hence in terms of the standards of achievement demonstrated by students. Standards of achievement are in principle indicated by whatever grading system is used for evaluation. Herein lies the problem, because standards of achievement are confounded with standards of evaluation. It is disentangling these two variables that poses the greatest challenge of all for the establishment of valid equivalence judgments.

Evaluation can take many forms, and final grading judgments can be made in several different ways. Assessment can be continuous or instantaneous, formal or informal, involving standardised, national or teacher-made tests, or even no tests at all. Where examinations are involved they can be single-component or multi-component, they can be entirely paper-based or can involve practical or verbal production, and so on. Student performances might be judged by a single individual or by more than one individual, using a detailed analytic mark scheme or by global impression, on their own merits or through comparison with others. Different components might be evaluated by different individuals using different approaches, and then their separate results might be weighted before aggregation. The criteria for classification might be determined in advance, based on judgments about appropriate levels of achievement, or after the event in the light of performance information. And finally, the classification system might be letter-based, number-based or even word-based, with any number of classifications possible.

Without complicating the matter further with any discussion around the rather important topic of evaluation reliability, how can we set about making decisions about equivalences between grading scales in the light of this diversity of practice? Is there any simple way in which correspondences might be drawn between the points on one nation's grading scheme and those on that of another? Unfortunately, there is not.

The American education system universally uses the grading scale A, B, C, D and F, with the possibility of pluses and minuses at the choice of individual teachers or institutions. It is generally assumed that the grades translate verbally into the concepts 'excellent', 'above average', 'satisfactory', unsatisfactory' and 'fail'. In the UK different scales are used at the levels of secondary and of tertiary education. The pre-university A-level system uses a scale running from A to E (minimum passing grade), plus F (fail). There is also an O grade, which concedes a pass at the old General Certificate of Education 'Ordinary Level' (examination usually attempted at the end of the compulsory period of secondary education two years earlier). Within higher education the usual classification scale is First Class, Second Class (Division I and Division II), Third Class, Pass and Fail. In neither sector is there any claim that the classifications correspond in any direct way with qualitative verbal statements conveying a sense of absolute meaning. In Switzerland different scales are used in different parts of the country, in secondary and in higher education, and again there is no ready verbal translation more meaningful than 'highest', 'next highest', and so on. The most frequently used scale is a six-point scheme, ranging from 1 (lowest) to 6 (highest), with a pass mark of 4. An alternative six-point scheme runs the other way, *viz.* from 6 (lowest) to 1 (highest), with a pass mark of 3. The third, 10-point, scale runs from 1 (lowest) to 10 (highest), with a pass mark of 6.

These three countries alone exhibit variation in their patterns and policies for grading. Other countries in the world, within and outside of Europe, merely add to the diversity. How then do we cope with this pattern of difference in any attempt to make statements about equivalence?

The ages at which the evaluations are conducted tell us something about relative standards of achievement. On a purely maturation argument, we assume that there is some natural degree of equivalence in standards of achievement among the various pre-university certificates awarded in Europe; more correspondence at least than would be expected between secondary and tertiary certificates within the same country. But would we be justified in drawing simplistic direct correspondences between grading scales, and inferring anything valid about equivalence? Can we, for example, infer that a grade D at British A-level is equivalent to a 4 on the Swiss Maturità? Possibly we can. But we might be wrong in doing so. And if they are equivalent, are they in turn equivalent to a grade D in the US college system? Or to a C, or a B, or even an A? Or to none of these? How do we judge? And what happens when we bring into the picture the French Baccalauréat, the German Abitur, the Dutch VWO, the European Baccalaureate, the International Baccalaureate, all the other pre-university certificates of Western Europe, and then all the corresponding certificates of Eastern Europe with which we have even less shared familiarity?

To help resolve the problem we might resort to entry and pass rate statistics. For suppose that in one country 80% of the age-group attempts

the relevant examinations and the majority of these candidates pass. In another country it might be that only 20% attempt the examinations and barely half of these pass. If we could happily accept the notion that the two national populations were inherently similar in terms of intellectual calibre, then these statistics would tell us something about relative standards of grading, and hence of achievement. The majority of certificate holders from the one country could be assumed unlikely to be capable of gaining the certificate of the other.

An appeal to such statistics would at least discourage the kinds of naive and totally unjustifiable statements about grade correspondence included in the less responsible American guideline publications. But even with such statistics to hand, and interestingly they are rarely recorded in comparative accounts, we still have problems determining appropriate correspondences amongst grades. Does it follow, for instance, that anyone holding a certificate from the second country, whatever the finer grade classification, would be equivalent in terms of past standards of achievement to those certificate holders with the highest possible grade in the first? What exactly would be the appropriate correspondence?

Grade comparability investigation is fraught with difficulties, and in consequence offers the greatest educational and technical challenges. The UK is arguably the country with the longest experience in this field. At pre-university level the UK examination system revolves around several relatively autonomous examining boards, which prepare examination papers and grade performances in various subjects. The grading scale used is common to all boards, with the consequent invitation to assume grade equivalence between boards in terms of relative standards of achievement in particular subjects. Anxious to validate this assumption, the boards themselves and individual researchers embarked in the mid-1960's on what became a very long programme of grade comparability investigation, focusing on the external examinations held at the end of compulsory schooling as well as those at A-level. Reviews are available of this work (Bardell, G. S., Forrest, G. M., Shoesmith, D. J., 1978, Johnson, S., Cohen, L., 1983, Forrest, G. M., & Shoesmith, D. J., 1985).

The British experience has relevance for Europe to the extent that these were attempts to establish grade equivalences among examining institutions whose syllabuses evidenced differential subject coverage, and whose examination structures, evaluation policies and practices also varied. Indeed, it is probably safe to say that no other country has for so long resisted the temptation to limit diversity, whilst at the same time attempting to research the consequences in equivalence terms. Europe is in that situation at the present time.

All the usual statistical strategies were first tried in the British comparability work, including the use of common components and of 'reference tests', but with unconvincing results. This initial exploratory phase was followed by a series of investigations based on the more expensive and more time-consuming approach of 'cross-moderation', in which examiners themselves were required to review candidates' examination performances and to produce comparative judgments about relative grading standards. This approach showed more promise, but its implementation was flawed in several ways, not least through an

8.1

assumption that a single subject examiner could 'carry' and transfer institutional standards. A formative evaluation of the cross-moderation method showed that, with an appropriate experimental design underpinning the deployment of the human judges and the analysis of their grading judgments, this approach could produce interpretable results about relative standards of grading, at least in availability of performance evidence and circumstances of high curriculum equivalence (Johnson, S., Cohen, L., 1983, Johnson, S., Cohen, L., 1984, Johnson, S., 1989). A combination of quantitative and qualitative approaches proved the best way forward.

Another finding of relevance for this chapter is that the evaluation study which focused on A-level mathematics revealed clear evidence of differences in grading standards amongst the three boards which participated. The differences varied in size at different grade levels, but were large enough that thousands of the holders of an A-level certificate in mathematics from one board would have been awarded a higher grade had they attempted another board's examination at this level, and vice versa. One can only contemplate the many injustices suffered by individuals who lost the chance of a university place on the strength of their A-level results, or the unfair advantages inadvertently offered to those who gained at their expense.

Within higher education in the UK, institutions have traditionally enjoyed a high level of autonomy. As in most countries of the world, individual teachers in this sector evaluate their own students, usually with their own evaluation methods, materials and grading criteria. Evaluation standards are in principle monitored through a system of 'external examining', or 'moderation' as it is more generally termed in other contexts. No empirical studies have been conducted into the issue at this level, but the implication from secondary-level research must be that differences do exist in grading standards between institutions (Johnson, S., 1988). Differences are indeed acknowledged by those within the system, but these remain to be quantified.

Switzerland imposes curriculum standardisation on schools for the Maturità, and has an external system of examinations for the Federal version of this certificate with pre-determined pass marks. The system appears to be controlled. Yet what do the various grades actually signify? What are the appropriate correspondences between the different grading scales? And are grades comparable from one year to another in terms of underlying levels of student achievement? Standardisation of subject coverage and examining policy does not necessarily guarantee grade equivalence. If we cannot comment on grade equivalence in pre-university certificates, what can we say for higher education? As in the UK, Swiss institutions of higher education have a high level of autonomy in curriculum structure and evaluation practice (Jablonska-Skinder, H., Teichler, U., 1989), with clear implications for certificate equivalence.

In America there is in general little or no standardisation of curriculum or evaluation practice at school or college level, other than through the use of the common grading scale. There is certainly no control over grading standards in either sector. Hundreds of thousands of individual teachers evaluate their own students according to their own criteria and personal standards. Most of those working within the system would readily

acknowledge that grade comparability does not exist; those few attempts to investigate the possibility empirically would confirm this to be the case (Strenta, A. C., Elliott, R., 1987). For suppose a student is considered to have achieved to a 'satisfactory' level in a particular course, and hence is awarded a grade C. What does this mean? Satisfactory in relation to what? What has the student actually achieved? How might that student have been graded by a different teacher, in the same or a different institution, at the same or a different time, or even by the same teacher in another semester? We do not know.

It is precisely because of this lack of grade equivalence that US colleges and employers make such extensive use of standardised 'predictor' tests when selecting students and employees; tests such as the Scholastic Achievement Test (SAT), the Graduate Record Examination (GRE) or the Graduate Management Achievement Test (GMAT). The absence of quality control of output leads to the need for quality control of input. This might meet the needs of the system adequately enough. But does it do justice to the individual? And can we defend the practice on educational grounds? These questions have been raised many times before, and are still in debate.

If we cannot ensure grade equivalence within the educational system of one nation, then how can we make any useful statements about grade correspondences across frontiers? Much remains to be learned about the intellectual processes involved in grading academic performance, and about the mechanics underpinning grading practice in different countries, before we can move forward on this front. Until we increase our knowledge, and improve our understanding of the fundamentals involved, progress towards solutions in this particular endeavour will be slow.

In conclusion

Europe has some difficult choices to make for the future, if a valid system of academic recognition is indeed to be implemented for European certificates.

One way to solve the equivalence problem would be to constrain the present diversity within and among the different national educational systems. But then we must decide how to do this – which features of which existing system should be retained, and which eliminated, in order to come to some acceptable degree of standardisation? The ECTS initiative will undoubtedly result in some bilateral and even multilateral agreements on standardisation among individual institutions, at least as far as curriculum is concerned. But the national introspections and cross-national negotiations which would be required for change on a larger scale, affecting evaluation practice as well as curriculum development, would take time, and meanwhile the issue of equivalence would persist.

Alternatively, one could choose to live with the current diversity, and to find some kind of workable solution to the problem of equivalence which would serve the system as a whole whilst at the same time minimising injustices to individuals. We could attempt to identify more finely those courses seeming to share a high percentage of common ground in terms of

subject coverage. We could amongst all courses offered across Europe make consensus decisions about the degree of equivalence which could be assumed in different contexts, using as much descriptive information as possible. We could begin by requiring that all teachers and examining bodies in every European country produce course descriptions, including objectives and content syllabuses, and that they also make available copies of their examination papers and detailed accounts of their grading procedures.

If we wish to tackle in a serious way the question of equivalence in terms of standards of achievement, accepting the current situation of extreme diversity in practice, then there is only one route to take. This would require evaluators themselves, teachers and examiners, to make comparative judgments, actively scrutinising the demands made on candidates in different examinations and the candidates' resulting performances. On the basis of all the cumulated experience in this field over the past three decades, 'cross-moderation' appears to be the only strategy which can produce interpretable results, and only then if applied in studies based on careful experimental design (Cohen, L., Johnson, S., 1982). The approach is time-consuming and logistically challenging, particularly given the additional problems posed by language differences within Europe. For all these reasons it would be expensive. And its success rate in allowing unchallengeable inferences about grade comparability would undoubtedly vary by subject, since courses involving oral or practical components in their evaluations might have to be judged on partial performance evidence after the event. Despite these drawbacks, if valid statements about grade equivalence are needed then this is the way to go.

If Europe is not prepared to embark on a grade comparability exercise on such a grand scale, first within each country before contemplating extending the programme across frontiers, what then? Certainly, any attempt to make 'educated guesses' about grade correspondences, however sophisticated, should be avoided. The best we can do perhaps is to provide the kinds of entry and success rate statistics referred to earlier. At worst, we might consider the large-scale deployment of 'predictor tests', as used in the US. If we go down this road, then let us at the least produce European tests, whose risk of culture bias will be confined to the European arena. And let us worry long and hard about the equivalence problems generally associated with language translation.

Perhaps it is inevitable that published guidelines will leave much to the discretion of those individuals making decisions about admission or academic advancement. But surely the more detailed the documents, the more information they include and the more wide-ranging this is, the more intelligently might those decision-makers come to their conclusions about appropriate action. Information is needed, about systems, about courses, about pass rates, about classification scales and, most importantly, about evaluation practice. It is entry and pass rate statistics, and information about evaluation practice, which are currently the missing elements in otherwise comprehensive documents such as Unesco's Higher Education Diplomas in Europe (Jablonska-Skinder, H., Teichler, U., 1989). Yet until we have the complete picture we cannot know the true extent of similarity and difference between national systems of education. And without complete comparative

knowledge we cannot make judgments about the potential for identifying equivalences, nor about the degree to which some standardisation might be useful.

References

AACRAO, 1988/1989, *International Academic Credentials Handbook*, Volumes 1 & 2, American Association of Collegiate Registrars and Admissions Officers.

Bardell, G.S., Forrest, G.M., Shoesmith, D.J., 1978, *Comparability in GCE: A Review of the Boards' Studies, 1964-1977*, Manchester, Joint Matriculation Board.

Cohen, L., Johnson, S., 1982, *The Generalizability of Cross-Moderation*, in: British Educational Research Journal, 8, pp. 147-158.

Commission of the European Communities, 1989a, *Guide to the European Community Programmes in the Fields of Education*, Training and Youth, Brussels, Commission of the European Communities.

Commission of the European Communities, 1989b, *Report on the Experience Acquired in the Application of the ERASMUS Programme 1987-1989*, Brussels, Commission of the European Communities.

Commission of the European Communities, 1990, *ECTS, European Community Course Credit Transfer Scheme*, Brussels, ERASMUS Bureau.

ECIS, 1990, *The 1990 International Schools Directory*, Petersfield, UK, European Council of International Schools.

Forrest, G.M., Shoesmith, D.J., *1985, A Second Review of GCE Comparability Studies*, Manchester, Joint Matriculation Board.

Jablonska-Skinder, H., Teichler, U., 1989, *Higher Education Diplomas in Europe*, Brussels, European Centre for Higher Education (CEPES).

Johnson, S., 1988, *Comparability in Degree Awards: Implications of Two Decades of Secondary-level examinations Research*, in: Studies in Higher Education, 13, pp. 177-187.

Johnson, S., 1989, *Évaluation de la comparabilité des notations entre jurys d'examen*, in: Mesure et évaluation en éducation, 12, pp. 5-21.

Johnson, S., Cohen, L., 1983, *Investigating Grade Comparability Through Cross-Moderation*, London, Schools Council.

Johnson, S., Cohen, L., 1984, *Cross-Moderation: A Useful Comparative Technique?*, in: British Educational Research Journal, 10, pp. 89-97.

NAFSA, 1984, *Handbook on the Placement of Foreign Graduate Students*, Washington, DC, National Association for Foreign Student Affairs.

Nisbet, C., 1980, *Advanced Standing*, NAFSA Newsletter (October).

Sepmeyer, I.H. (ed.), 1986, *The Country Index*, Revised Edition, Vol. I, North Hollywood, CA, International Education Research Foundation, Inc.

Strenta, A.C., Elliott, R., 1987, *Differential Grading Standards Revisited*, in: Journal of Educational Measurement, 24, pp. 281-291.

10 Examinations in Europe: Three Major Streams

H. P. J. Kreeft

Introduction

The systems of evaluation and examination vary considerably in the member states of the European Community. No two countries have the same system. With respect to final assessment, each country has its own history and development, running more or less parallel to the development of the educational system. Some countries, such as France and Ireland, have a strong external examination system. In other countries final assessment is largely a responsibility of the schools themselves. This is called internal assessment. Examples are Belgium and Spain. In between these extremes, any combination of external and internal assessment may occur. Examples are the Netherlands, Germany, and the United Kingdom.

It has been said that the advantage of external assessment is its recognizability. It is clearer to the receiving educational institutions what may be expected of the student. Trade and industry benefit from a certain degree of uniformity in reporting and certification. Internal assessment might raise doubts as to the value and validity of the certificates. External assessments provide the authorities with an effective means of maintaining standards.

On the other hand, internal assessment has been said to have the advantage of creating the opportunity for not only assessing the learning outcomes, but also the way in which they were produced. The chances of failing by accident are smaller, since the assessment may take place more frequently and in more varied forms. Another advantage which is often mentioned is that the backwash effect, which occurs with external assessment, is absent in the case of internal assessment. Whether this latter argument should always be considered an advantage, is not quite clear to me. Internal assessment also has advantages of a different kind, namely in terms of organisation and finance.

In deciding whether external assessment should be replaced partly or entirely by internal assessment, these considerations will not be the least important. Much more can be said about the pros and cons of external or internal assessment. In this chapter, however, I would like to concentrate on the aspect of transparency for continuing education, and trade and industry.

Problems with the information value of certificates may occur within a country. When the diversity of certificates and final assessments becomes too great and too subtle, complaints develop. This is the reason, for example, why the GCSE was introduced in England. When, however, the exchange of students and labourers between the member states of the EC increases after the unification of Europe, the transparency of certificates will be of great importance in this context as well. The need for maintaining standards, which is common to all countries, is closely linked to this.

In this chapter, I will try to give an idea of the degree to which different countries have external or internal assessment, with a view to the transparency of the certificates and with a view to the possibilities for maintaining standards. I would like to review the developments in the past and the current development. Furthermore, I would like to apply considerations of test quality to the context of external versus internal assessment, and to transparency of certification.

External centralized examination systems

Examination systems may be called external and centralized when exams are organized and administered by a body outside the school. The French 'baccalauréat' is a good example of this. In the French system, the exam regulations have been laid down minutely by the ministry of education. Exam contents are determined by the 'Academie', as are exam questions. Throughout the country, exams are administered at the same time. Exams are not held in schools, but in exam centres. The rating is carried out by anonymous, specially recruited raters. Oral exams are administered by two examiners unknown to the candidate. A jury decides whether a candidate has actually passed or failed the exam as a whole. Candidates and schools are informed of the results after some time.

The question is whether such an external examination system has a maximum of transparency with respect to certification. In the case of the French 'baccalauréat' exams, quite a lot can be said against it. Of minor importance in this respect is the fact that, for organisational reasons, the exam questions are not identical throughout the country. France has some 30 academies that draw up different series of questions. The academies operate in clusters and the questions are based on the same regulations issued by the ministry of education.

Much more detrimental to the transparency and to the degree of centrality, however, is the type of question that is often used; the so called 'dissertation' and the commentary. It consists of one single question or one set of material. The candidate is completely free in deciding how he responds to it. He roughly knows from the preparation for the exam, however, what is expected of him. With such a task, the candidate himself determines what knowledge and skills to demonstrate. The task is not the same for each candidate.

In an external, centralized examination system, the actual approach that is taken may seriously affect the centrality and, consequently, the transparency. The French way of constructing questions is also problematic as far as maintaining standards is concerned. Standards are maintained in the sense that the exams are selective, but what the exact standards are is difficult to define for many subjects.

Internal examination systems

The other end of the continuum from external to internal is formed by examination systems which are entirely internal. Internal or school-based assessment is a form of final assessment in which the exams or test items are constructed by the teachers of the school itself. The items are not judged beforehand by any external body. The administration takes place within the school at points in time determined by the school. The rating is carried out by teachers of the candidates' own school. Decisions on pass or fail are taken by the staff meeting. Examples of such a school-based system can be found in Belgium, Greece and Spain.

The freedom of the schools with respect to final assessment seems maximal. Yet, in Belgium for instance, considerations of quality control have led to some measures that reduce the school-internal character. In the new Belgian system of secondary education, the two final grades form the determination years, in which the summative evaluation has to take place on the basis of central curricula. The adequate implementation of the curriculum is monitored by the inspectorate. In addition, Belgium has a so-called homologation committee, which makes random checks on the evaluation by the schools afterwards.

The transparency of the certificates issued is supposed to be particularly promoted by the curricula. That the transparency is in fact too low, due to the diversity of interpretations given to the curricula in the permanent evaluations, is apparently not perceived as a problem in Belgium.

In Spain, the low transparency of the final assessment has led to entrance exams for universities. Such entrance exams may be regarded as constituting an external exam, which lacks centrality, by the way, since it is organized by each individual university or department. State control is much less direct due to the absence of central, external papers.

Mixed systems

In the continuum defined by the two extremes there are various examination systems with external as well as internal characteristics. I will discuss the Dutch system as an example here. In this system, each school subject is tested externally once and a few times internally. Both components each contribute 50% to the final mark. The internal assessment takes place at different points in the year, a kind of continuous assessment. The external assessment consists of written nationwide exams, that are administered to all candidates at the same time during two weeks in May. In contrast to the external system in France, the external part is held in the schools. The rating is carried out by two teachers: the candidates' own teacher and a teacher from another school. The rating is neither anonymous, nor independent. Since the candidates' own teacher is the main judge, the formal proportion of 50% internal and 50% external assessment actually shifts in the direction of internal assessment.

The exam contents of both the internal and external parts have been laid down in syllabuses issued by the ministry of education. However, that does not alter the fact that the actual way of examining may strongly differ from school to school in the internal part of the exam. In matter of fact, this is exactly what the system is aimed at. The internal part should enable schools to evaluate the school-specific objectives. Debates in the Netherlands about no longer conflating the internal and the external assessments in the final decision, point to problems with the information value of the certificate.

Another example of a mixed system is the final assessment in the Italian 'lycea' (senior secondary education). Here, all subjects are evaluated by the schools themselves by means of a form of permanent evaluation. Candidates that have passed this school-internal part are admitted to the so-called Maturity Exam, in which 4 subjects are tested by an external committee, 2 of them orally and 2 in writing. In this external exam the same kind of questions are used as in France, which reduces again the centrality and transparency of the exams.

The central parts of mixed systems may provide an indication of the educational standard, provided the nature of the questions allows this. With completely open-ended questions, the question is whether the information value is high enough for drawing conclusions about the level tested. That the central, external questions in mixed systems usually do not cover the entire domain of objectives, or that not all subjects are examined externally, is a limitation of the exams in maintaining standards.

So much for the examples of the three different streams in the European examination landscape. It would carry us too far to discuss the systems of all countries here. What I do want to do however, is to locate the systems of the different countries on the continuum from external to internal examination, and to provide some clarification where necessary.

Survey of the three examination streams in senior and junior secondary education

Figure 1 gives a survey of the three examination streams for senior secondary education in the countries of the EC.

Figure 1 Characterization of exam systems for senior secondary education

External assessment	Mixed systems			Internal assessment
100%	75/25%	50/50%	25/75%	100%
	UK			
France	Netherlands	FRG	FRG	Belgium
Ireland	Italy	south	north	Spain*
Luxembourg	Denmark			Portugal*
				Greece*

The United Kingdom occupies a special position in this diagram. All combinations of external and internal assessment occur, although the form in which the major part of the A-level exams is administered by the examination board appears to predominate. Spain, Portugal and Greece have only internal assessment at the end of senior secondary education. But for admission to higher education entrance examinations are organised. The entrance examinations take the place of the final examination at the end of secondary education. Those are mostly external examinations, however, the results in the last years of the secondary education are also taken in account.

The same diagram can be made for junior secondary education at the end of compulsory education. We can see, then, that the emphasis has shifted towards internal assessment in a number of countries. We can see that the French, Luxembourg and Italian systems shift in the direction of internal assessment. Apparently, the function that the external assessment fulfils in the higher-level examination is considered less important here.

Figure 2 Characterization of exam systems for junior secondary education

External assessment	Mixed systems			Internal assessment
100%	75/25%	50/50%	25/75%	100%
	UK			
Ireland	Netherlands Denmark France	FRG south	FRG north	Belgium Luxembourg Italy Spain Portugal Greece

Shifts

Examination systems are subject to change, of course. Thus, the preference for external, internal or mixed systems is not constant either. Changes have usually taken place in the direction of internal assessment. Spain and Greece, for instance, had external final assessment in both senior and junior secondary education. These systems have been abolished in both countries. One of the main reasons was that the examination system was identified with the totalitarian political system that had come to an end and that was biassed in favour of the socially privileged. The reduction of state control involved the simultaneous abolishment of the external examination. In France we saw that in 1977 the external examination in the 'collège' (junior secondary education) was replaced by a system of permanent and, therefore, internal evaluation. In 1986 the present mixed system was introduced. The step from completely external to completely internal had been too big.

Currently, we can also see movements in the direction of internal examination. In Ireland, which has a virtually completely external system, there are experiments with practical exam components that will be administered by the schools. In France there have been discussions for some years now about introducing an internal component in the 'baccalauréat' exams. One of the reasons why the educational unions are opposed to this is that a school-internal exam would disturb the relationship between teacher and students, or that teachers would come under pressure from the parents to give higher marks to the students. Examination forms being so little objective, this fear is not entirely unjustified, I think.

The Netherlands are working on a common core curriculum for junior secondary education. Final assessment at the age of 15 to 16 will probably become entirely internal. A shift, therefore, from a mixed system to an internal system.

There is also a growing interest in school-internal assessment in England. Preparations are currently made for the introduction of a National Curriculum with test rounds at the ages of 7, 11, 14, and 16. The tests to be administered in this framework will be school-internal to a larger degree than was the case until now. Records of achievement, which will replace the exam certificates in a number of schools, also place more emphasis on school-internal evaluation. It should be added, however, that internal assessment in England has a different character than in many other countries. Internal assessment is intensively supervised by the examination boards. Objectives have been defined; test components are made available; assessment methods are practised with teachers. On the other hand, concurrently with this shift towards internal assessment, the influence of the authorities on the educational objectives and on the boards is being strengthened. For the GCSE General Certificate of Secondary Education (junior), National Criteria and Subject Criteria were determined. The development of the National Curriculum strengthens this influence even more.

Replacing external assessment by internal assessment requires more elaborate regulation, at least if a government wishes to have control over the educational standards and wishes to promote transparency. A similar link was made in the OESO review of Dutch education that was recently carried out. The investigators argue that the Dutch policy of liberalizing the current regulation and supervision, only increases the need for adequate quality control and assessment.

To summarize: countries with a completely external system of examination form a minority by now. Wherever shifts do occur, they go towards introducing more internal elements in the examination. The reasons behind this are of an educational nature. Changes in the educational system entail changes in the examination system. Apart from these educational reasons, organisational and financial reasons always play a role in decisions of this kind as well. Some forms of evaluation are very difficult to organise in an external manner, and only at very high costs.

Quality requirements of examination

A final exam that is of great importance to the candidate, that should enable the authorities to maintain standards, and of which society is entitled to demand that it leads to transparent certification, must meet high standards of test quality. I will distinguish three such standards in that a final exam must:

- have high content validity;
- be objective;
- be communicable or transparent.

Content validity

Content validity concerns the degree to which exam questions adequately represent the examination programme, or the final objectives as described in the curriculum. Exam constructors have the obligation to indicate in what way this representativeness has been pursued. Two situations: in England, the ministry has set National Criteria and Subject Criteria for the GCSE exam (junior secondary education). On their basis, the Examination Boards compile the syllabuses that make clear to the schools what the final objectives are and what the subject matter for the exams is. The syllabuses are checked by the Secondary Examination and Assessment Council, an organisation closely allied to the Ministry of Education. Exam questions are formulated on the basis of these syllabuses. I think this procedure is a maximum guarantee for content validity of exam questions.

In the German state of Nordrhein-Westfalen, exam questions for the so-called Abitur exam (senior secondary education) are formulated by each teacher for his own candidates. He does so on the basis of a curriculum determined by the ministry and guidelines (the so-called 'Richtlinien') which are issued by the Ministry of Education in the state of Nordrhein-Westfalen for each of the school subjects. The draft questions are subsequently sent to the inspectorate. The questions must be accompanied by a description of the actual education during the last two years and a description of the objectives tested. Therefore, content validity is being related the guidelines and the curriculum on the one hand, and the specific teaching situation in a particular class on the other. A lot of attention, too, in this case, for precision and accountability.

Yet, I think that these two procedures ought to be judged differently as regards content validity and communicability of exams. Firstly, the operationalisation of the English National and Subject Criteria takes place five times in the five regional clusters of cooperating Boards; in Nordrhein-Westfalen it takes place at least 600 times at each separate school. In England, exam specialists work on the questions, in Nordrhein-Westfalen it is one of the many duties of the teacher. I think this has consequences for content validity. It definitely has consequences for the communicability and transparency. Although the final objectives have been formulated in Nordrhein-Westfalen, and although they are accessible for anyone, the certainty that these final objectives will actually be tested is much smaller. There is another difference between the situation in England and in

Germany. The question formats used differ considerably, and this also has consequences for content validity and for the quality of the test.

In the United Kingdom, the candidates are presented with an extensive series of questions in different sessions. For each question it can be indicated which objective is being tested. In Nordrhein-Westfalen, a form is frequently chosen in which the exam consists of one question or a few questions. The relation between this question and the different final objectives can only be indicated very generally. After all, it is left up to the candidate to decide how he answers the question. The candidate himself decides which concrete objectives he will focus on in demonstrating his skills. With this, we have lost control over content validity. I have pointed out similar problems in France earlier on in my presentation. The content validity of an exam is ensured best in an external, central exam in which a sufficient number of questions are used. only this approach enables one to communicate with the outside world about the teaching objectives being tested; only in this way transparency is achieved.

Objectivity

The second quality requirement that exams should meet is that the scores of the candidates are determined in an objective manner. Objectivity is related to question types and rating methods. I have already mentioned the essay questions used in the German exams. In France and Italy too, frequent use is made of unstructured essay assignments. Apart from the fact that it is impossible to determine the content validity of such questions, they can also be hardly assessed objectively. Most of you will be aware that, if different judges assess the same work, very large differences may occur with this type of questions. As a consequence of the large freedom given to the candidate in answering the questions, it is also impossible to formulate assessment models. Giving a number of general criteria is the best one can do.

That one is aware of these problems, may be deduced from the fact that the exams in France are assessed anonymously, and that the jury of the Academy, which determines the final results, checks the scores given by the raters for extremely high and low averages. A measure that would be more effective – to have each exam assessed by a second or third rater – is not taken, I have been told, for financial reasons.

In Nordrhein-Westfalen there are two ratings: the candidates' own teacher makes the first assessment, while a colleague from the same school sees to it that he can agree with the assessment. Independence and anonymity are not guaranteed here. The assessment of questions that leave the candidates less freedom, such as questions that may be answered by a limited number of lines of text, or short-answer questions, may also be deficient in terms of objectivity. Differences between scores across an entire exam may easily run up to 10-20%.

In view of this, it is amazing that so few countries have a second rating. Only in the Netherlands, Italy, Denmark and in Germany do we find such a double rating. Even in England, where a great deal of attention is paid to the rating, most exams are rated only once. One tries to solve the problem there

by having so-called moderating meetings, where raters exchange experiences and come to agreements about the rating method.

What the situation is like in the internal examinations in the case of school-based assessment, is not entirely clear. But I guess I'm not being unfair when I assume that multiple rating is not very frequent there. In addition self-constructed assignments and questions will rarely be accompanied by detailed correction models.

The pursuit of a high level of objectivity is best realized in external, central examination systems. I have already pointed out, however, that this is not always the case in externally organized exams.

The lack of objectivity is not only problematic for candidates that fail on questionable grounds; but it also reduces the transparency of exam results, since it remains unclear whether good results are directly related to adequate mastery of the objectives or to a positively biassed assessor. This lack of informational value also affects the process of maintaining standards.

Concluding remarks

In summary, I think the following observations are warranted:

- that in Europe the emphasis is on internal assessment;
- that internal assessment is more frequent at the end of junior secondary education than it is at the end of senior secondary education;
- that changes in the examination systems involve a shift towards internal assessment, and
- that quality criteria in terms of content validity and objectivity, regardless of whether the assessment is internal or external in nature, are often not met.

From the point of view of recognizability and transparency of certificates, this is a problematic situation. The lack of transparency also constitutes a problem in terms of quality content in education. The loss of opportunities for maintaining standards is certainly not always compensated for by more regulations regarding teaching content and testing. The best opportunities for meeting the quality criteria mentioned and for maintaining the communicability of the final objectives in education, I think, lie in central-external examination. If there are educational and organisational reasons for part of the final exam being an internal evaluation, and there are such reasons indeed, then it seems to me desirable not to let them become too prominent.

A great deal of effort is required in order to make internal examination meet the quality requirements we have discussed. Regulations regarding testing and teaching form part of such an effort. Teachers have to be able to use tests or test models developed by experts. A form of external supervision and control is required for the administration, the assessment, and the translation of test results into final judgements.

From the point of view of recognizability of certificates in a future Europe, it would not be a bad thing if the three streams present at the moment would converge somewhat in the direction of mixed systems with an emphasis on external examination and external supervision. A prerequisite would be, however, for the external examination to meet high quality requirements.

References

British Council, The, 1988, *International Guide to Qualifications in Education, 2nd ed.*, London, Mansell.

Eurycide, 1986, *Der Aufbau des Bildungswesens in den einzelnen Mitgliederstaaten der Europäischen Gemeinschaft*, Brussel, Eurycide.

Gordon, J, 1990, *Comparative Study of Qualifications at the End of Compulsory Education, Secondary Education of Vocational Training*, Volume 1, Synthesis Report, Paris.

Postlethwaite, T.N. (Ed.), 1988, *The Encyclopedia of Comparative Education and National Systems of Education*, Oxford, Pergamon Press.

11 The Politics of Assessment

D. L. Ongom

Introduction

Politics may be defined as the art and science of governance. It refers to the exercise of power of man on his fellow men to influence the society or the state. Power may be seen as the ability to influence the behaviour of A by B so that A behaves in a way he otherwise would not have behaved if B had not acted. Politics recognises that in the exercise of power by man on his fellow men, the wielder of power seeks to influence the acquisition and allocation of material resources as well as controlling intellectual resources in society (Newmann, F., 1950). Morgenthau points out that the hallmark of politics as long as it involves someone seeking to exercise power over other men makes politics evil, and that all subsequent options in politics are evil, and therefore the power wielder has the unfortunate duty of choosing the lesser evil among the evils available to him (Morgenthau, H., 1945).

In theory, politics seeks to solve problems affecting the public in a systematic and orderly manner. In practice, however, some sections of the society may employ politics to influence the citizens in their favour. Success in politics depends on the ability of the individual to persuade and win the majority to a particular point of view. At times, this may be achieved through means which may not be considered honest or fair. To 'assess' means to gather information about the individual, process the data and then inform the relevant authority about the individual. Anderson, Ball, Murphy et al. defined 'assessment' as the process of gathering the data and fashioning them into an interpretable form. Judgement can then be made on the basis of this assessment (Anderson, S. B., Ball, S., Murphy, R. T., et al., 1976).

In general terms 'assessment' may be considered as all methods customarily used to appraise performance of individual pupils. It may also refer to a broad appraisal including many sources of evidence and many aspects of a pupil's knowledge, understanding, skills and potentials. 'Education assessment' aims at providing accurate and relevant information about the progress of the pupils, the value and relevance of the curriculum and effectiveness of teaching. Frith and Macintosh defined education assessment as "measurement of the extent of learning in the individual" (Frith, D. S., Macintosh, H. G., 1984, p. 248).

Assessment practices

In an educational system, assessment takes place all the time in the classroom. External assessments conducted by assessment institutions

occur at defined nodal points. In the case of Uganda, every child attending formal education must face the Primary Leaving Examination (PLE) after seven years of primary education, the Uganda Certificate of Education (UCE) after four years of education in lower secondary, and then the Uganda Advanced Certificate of Education (UACE) after a further two years of upper secondary education. The children do not choose what they consider to be the appropriate time for them to be assessed. This is decided long in advance by the system of education, although the individual child may postpone the time of sitting by a year or so, otherwise they have to conform or withdraw altogether. The power of assessment system which schedules the time and the date of examinations with no consultation with its clients, namely the school administration and candidates, has led to some amount of discontent.

Anyone taking part in formal education is a candidate for assessment (Broadfoot, P., 1979). Normally this are children attending school. Apart from the children themselves, their parents and their teachers too, become candidates on behalf of their children and students. Parents view the failure of their children in the examinations as their own failure. Teachers also view the failure of the students they have taught as their personal failure. The feelings pupils may have against public examinations in particular and assessment practices in general are equally shared by both their parents and their teachers.

Assessment practices are based on the objectives and contents of the syllabuses prepared by the curriculum developers. The curriculum developers should clearly state the regulations that control assessment practices in an education system. However, in practice, due to the great influence and power of the outcome of an assessment, they tend to overemphasize what to teach and how it should be taught. Teaching is often distorted and only what is likely to come in the examination is taught. Broadfoot has put it quite correctly when she accused the "power of assessment tail to wag the curriculum dog and to impede the curriculum development..." (Broadfoot, P., 1979). In normal practice the examination should not dictate what should be taught, it should merely test what has been taught. The powers of assessment often offend the curriculum developers.

Assessment institutions in collaboration with the curriculum development institutions have devised what they consider to be efficient assessment systems. There is a serious debate about the use of public examinations versus school-based assessment; which assessment instrument to use; essay questions versus objective questions, and so on. Broadfoot summarises the discussion thus: "The focus of such discussions is overwhelmingly on examining the current practice and working out how it may be changed to become more efficient and thus ostensibly more just" (Broadfoot, P., 1979). The failure by psychometricians to agree on the standard method of assessment that is most efficient and therefore, acceptable to all concerned, signals the contest of power of assessment.

The purposes of formal assessment are many. One of them is accountability in education. The cost of education is extremely high and represents a major item in the national budget. Education must account for itself to the society and the nation. The success of education in a school is

measured in terms of achievement by the students in public examinations. In the nineteenth century, the need for accountability was expressed in the 'payment by results' system in which schools were given money in return for bringing individual pupils up to the required standard in certain specified areas (Broadfoot, P., 1979).

Assessment results are used for certification. Those who perform well qualify for the award of certificates. The possession of a certificate means success in education at a given level of academic competence, which is usually equated to social and administrative competence. Certificates serve as social indicators of ability and potential of the holder. They pin upon the holder some kind of class. The certificate plays the 'gate-keeper' role by which they open and close doors for individuals to future life chances (Broadfoot, P., 1979). The gate-keeper role attracts hatred from those shut out by the assessment system.

Assessment results are used to select and admit individuals for particular training courses. Earlier, in a paper presented at the conference of IAEA European Zone in Scotland (29 August - 2 September 1988) on the principles of certification for secondary school candidates in Uganda, I stated that a selection decision represents an attempt by the authority concerned, to pick the students who will have the greatest benefit from the course of training. The principle on which selection is based is a recognition that past achievements often have good predictive validity for the future performance. This assertion can be seriously contested.

Assessment results may be used to evaluate the existing curriculum. Through such evaluation, a change in the methods of teaching, teaching materials and emphasis in the syllabus contents may be reviewed and possibly modified. The curriculum developers may challenge the examination contents and results as they relate to the curriculum.

The politics of assessment

In developing countries such as Uganda, public examination results ignore all other educational values acquired by children as part of the activities in schools, except academic achievements. Besides, assessment outcomes are used to brand or label children. Whatever label a child is branded remains with him for the rest of his life. Serious decisions with lasting consequences are based on the results of assessment regardless of the instrument used, its reliability and validity.

The practices of assessment have therefore met with sharp criticism from all possible corners. The students, parents, teachers, educational administrators and the curriculum developers, have over the years detested the boundless powers wielded by the assessment institutions. They would like to find out who should have a bigger say in the field of educational assessment. The public too, being aware of the colossal amount of money spent on education would want to be convinced that the society is getting good value for the money. The society tends to believe that the best way to judge performance of education is to find out the performance of candidates in public examinations. The students, their parents and teachers,

educational administrators, and the curriculum developers all work hard towards achievement of educational goals. They are concerned about the degree of success as determined by an assessment system. However, they have different views of the best methods of achieving the goals or how best to assess the success of educational goals. They have reached a point when they feel they are entitled to examining the assessment practices currently in use. Some of them have gone further. They have questioned the needs for assessment in education, the process of assessment and the quality of assessment. They dislike the over-emphasised functions of results of public examinations.

The contest of the powers of assessment has led to what may be termed the 'politics of assessment' which refers to all the conflicting activities directed to persuade the assessment institutions to change the existing system of assessment in favour of one of the interested groups. The contest of the powers of assessment goes beyond the concerned groups, namely the students, parents, teachers, educational administrators, curriculum developers and the assessment institutions. There are, for instance, the politicians. The politicians being the main determinants in the laws, economy, materials, and how intellectual resources are organised, tend to reflect rather negatively over the powers of assessment institutions.

The assessment institutions other than performing delegated responsibilities from the political top hierarchy become power centres in themselves. They have the power to test, administer and manage individuals when they surrender themselves to the examining authority. Many of those who surrender to be examined do not do so willingly. They do so under subdued protests. They have come to hate examinations and devised ugly names for them. Examinations have, the world over, been called names that imply evil; for example, 'animals', 'the devils', 'modern mephistopheles', etc. A cross-section of society view examination as evil. Some people have accused examinations of exposing their intellectual shortcomings to the public, equating it to the exposure of a person's poverty, bank account or the disease a person is suffering from (in public). The majority would prefer mass collective security from the ugly pangs of examination and set up a system where they are taught but not assessed or ranked.

The fact that examinations are now taken for granted and believed as a necessary evil by persons who, by their nature, are conformists, should not blind the assessment institutions that there is not a silent majority who are non-conformists. These oppose the principles and practices of assessment and its outcomes but are neither organised, nor brave enough to say 'no' and have no power to renounce examinations. It is in the light of these inhibited resentments towards examinations that the 'politics of assessment' becomes prominent.

Causes of the politics of assessment

In developing countries, success in public examinations is seen as a guarantee for soft life provided by white collar jobs with all its trappings and

dignity. It is viewed as a guarantee to a middle class occupation. Any person who does not perform well in public examination does not accept the responsibility for the poor performance. The assessment institution is blamed for employing assessment practices which have caused those particular candidates not to do well in the examination. The candidates who do not do as well as they and their parents expected resort to accusing the assessment institutions of all sorts of weaknesses imaginable. They hit at examining institutions at the first available opportunity.

The influence and powers exercised by the assessment institutions which they mercilessly use to persuade the students and their parents, teachers, and employers that public examinations and only public examinations have the divine power to discriminate between the academically clever and those not so clever, are hated. Any system that acquires and exercises too much power is hated. In developing countries where any system that is seen to have too much power is violently overturned so that a completely new system is tried, it is clear that assessment institutions in those countries are more prone to being severely attacked to the point of destruction. Assessment systems in developing countries should take note and make all efforts to share its powers with their clients and those other institutions that are legitimately concerned.

The award of certificates based on the academic achievement of candidates in public examinations has formally legitimised the present dominant position of assessment. It has firmly established the superior role of academic ability and relegated non-academic aspects of education. Abilities in games and sports, music, dance, and drama and other forms of art are not rewarded by the existing systems of education. And yet those who live more comfortably after schooling are: soccer players like Maradona and Pele, musicians like Michael Jackson and Miriam Makeba, dramatists like Sean Connery and Sophia Loren, tennis players like Boris Becker and Ivan Lendl, and renowned boxers like Muhammed Ali and Mike Tyson. The assessment system is seen to have miserably failed in such cases, to identify real and more viable potentials. The emphasis on intellectual achievement in academic subjects has had the backwash effects of changing the purpose of schooling from emphasizing leadership training and healthy personal attitudes to preparing children to pass examinations. It has distorted good educational practices. Even in its declared job of identifying academically able students it has not been wholly successful. Great inventor Thomas Alva Edison was not discovered by a public examination. An assessment system is a system seen by many as a 'pretender' that cannot handle the job it pretends to be an expert in.

In order to pass examinations, students learn to cram and reproduce facts rather than developing creative thinking. Oginga Odinga once accused colonial education in Kenya of 'widening the cramming departments of students instead of deepening their thinking faculties'. He saw this type of education as producing students who were 'prone to taking orders instead of issuing orders'.

Broadfoot pointed out that "there are links here with the importance of assessment as a means of social control, because the encouragement of passivity and acceptance of the prevailing knowledge, values and standards, and the restriction of entry to elite position to those willing to conform to

such definitions are vital to the process of social reproduction" (Broadfoot, P., 1979). Many educators believe that public examinations as perpetuated by assessment institutions are in part responsible for the lowering of educational standards, except that they see almost no appropriate alternatives.

The status and the morale of a school is very much dependent on how well its students performed in public examinations. The performance of a school in a public examination has become the only indicator of the quality of education the school is providing. The education instruments, the process and the policy the school practices throughout the year, which are the real quality of education indicators are conveniently overlooked. Teachers whose schools excel in public examinations walk with their heads up in the local community and in public places like pubs; while those whose schools do not perform so well, and these are in the majority, keep away from the public and quietly curse the assessment institutions for their sorry position. They quietly promise to square up with the examiners at the appropriate time.

Assessment institutions have learnt that in their trade they must make several regulations to be followed by all their clients. Although in framing those regulations – often they follow the dictum of legal practitioners – they end up producing rather stringent regulations which they expect all their clients to follow religiously. If those regulations were closely adhered to, life would be made much easier for the institutions themselves in doing their work. Unfortunately, following regulations demands attention to details, an ability which very few people possess. Besides, there are people in the society who believe they are above regulations and rules. In ordinary life regulations are frequently broken and quite often ignored by many organizations. In some cases regulations are bent to suit some certain individuals.

For some unknown reasons, assessment institutions believe that rules and regulations once produced must be followed judiciously. According to the codes and ethics governing the employees of assessment institutions, rules and regulations once produced cannot be manoeuvred to suit individuals irrespective of their positions in the society. The main principle of assessment is that all must be equal in the process of examination. This principle though attempting to uphold justice, affects important people in commerce, government, and politics. They feel jilted and disregarded. They promise to retaliate in 'hard currency' should situations present themselves.

Many educated people in society feel that they know everything about education including examination. The naked fact is that they are largely ignorant when it comes to examining in particular and assessment in general. Assessment is a dynamic discipline which is undergoing constant changes. Those who are ignorant about the many variables that must be taken into account when dealing with the examinations, unfortunately include highly placed persons in commerce and industry, politicians and civil servants, and at times employees of the Ministry of Education. They all pretend to know all about the operations of examination institutions. Many of them criticise examinations merely out of ignorance.

In developing countries, education is considered to be a very important investment for the young ones. However, very few vacancies are created in institutions of higher learning. To move from one segment of the education

ladder to the next, there is a vigorous competition for the few available vacancies. In the case of Uganda, at the end of secondary education, students sit for the UACE examinations. The UACE examinations are used for admission to Universities and institutions of higher learning. However, due to limited vacancies in the universities and tertiary institutions, many students who also deserve to be selected are left out. Such students and their parents are severely frustrated. The frustration and anger resulting from failure of the university to admit such children is unfortunately not directed at the university or the Ministry responsible for Education, but squarely placed on the shoulders of the assessment institutions. The examining authorities in Uganda and elsewhere are accused of all types of weaknesses by the frustrated students and their parents.

Participants in politics of assessment

The participants in the politics of assessment are many, but the following have been identified:

- Politicians: namely the members of Parliament and local council, who normally raise issues during their debates.
- Educational authorities: education officers, inspectors of schools, head teachers, classroom teachers and curriculum developers. They normally write letters of complaints directly to the examining authority raising objections to certain practices in examinations and offering their considered advice. Close scrutiny of their letters of complaints, in most cases, reveals that they are written based on emotion and not through love and fairness of the examination system.
- Clients: candidates, parents and teachers. These normally write sensational articles in the press under screaming headlines. The press prefers to print those letters which are sensational and likely to boost sales of newspapers.

All the people who lock heads with examining authorities lump themselves under the guise of the general public. This gives a very convenient cover under which the emotional criticisms gain credence.

Stand of assessment institution in the face of the politics of assessment

The assessment institutions are legal entities established to perform a definite job. The work of examining is carried out by institutions set up by acts of parliament with defined functions and powers as enshrined in the constitutions. The examining institutions must be operated strictly according to the constitutions. To assist in their operations, examining institutions prepare guidelines or regulations. As long as the institutions operate strictly according to the constitutions and follow the regulations and laid-down procedures they should not meet problems. Problems met can always be clearly explained.

Those involved with educational assessment are forced to solicit an understanding from the authorities concerned regarding the optimum conditions they require to do their jobs effectively. If such an understanding is attained, educational assessment should then be carried out free of political and partisan pressures. At the same time, the political pressures to get things done regardless of the existing capacity and proper procedures should be resisted by educational assessors to avoid down-grading the importance of educational assessment.

By the nature of their duties, examining institutions are involved in educational measurements. As a result of the assessment, the examining institutions make decisions which are pleasant to some people and unpleasant to others. Those adversely affected by assessment decisions feel unhappy. They are the ones who hurl unwarranted invectives at assessment practices. Assessment institutions should have enough shock absorbers to withstand such attacks.

With an increase in the cost of examinations, it is natural for the public to show interest and concern. Mehrens and Lehmann pointed out that "There are few incidents of valid concern by the public, however, in general all those who criticise examinations fail to understand much in the field of educational assessment and the result is that their concern are frequently invalid" (Mehrens, W. A., Lehmann, I. J., 1978, pp. 667-87). Educational measurement is a technical field and one cannot understand it, let alone criticise it intelligently without some mastery of its contents. Those who believe examination results are not valid should note that there is a considerable misunderstanding among laymen about the concept of validity of examinations.

Assessment institutions in developing countries where continuous assessment results are ignored must be cautioned. There is room for combining some elements of examination results and records from continuous assessment to develop a complete profile of the individual candidate. The assessment institutions should by now have been aware that most criticisms about examinations are due to ignorance on the part of the public about the methods of operations of assessment institutions and not on fundamental issues. Assessment institutions as professional institutes must not and cannot afford to be intimidated by irresponsible critics of their examination systems. They must, however, ensure that valid examination results are used for appropriate decision making.

Concluding Remarks

Politics is the art and science of governance. The wielder of power in politics seeks to influence the allocation of material resources and controlling intellectual resources. Assessment is the method of appraisal of performance of individual pupils. Assessment institutions particularly in developing countries do not take into account the aspirations of its clients and the co-operation of its partners. As a result the institution is viewed by the public as wielding too much power. Besides, assessment practices are viewed as harsh and insensitive to those directly affected by the assessment

outcomes, and as such, assessment practices have been subjected to much criticism from the public. This has led to some sections of the public to contest the powers of assessment. The contest of powers of assessment has resulted in the politics of assessment. The causes of the politics of assessment arise from two main factors; the effect of assessment results on its client and the uncompromising adherence to its regulations and procedures by assessment institutions. Assessment institutions are advised not to yield their ground in the face of politics of assessment. They should defend their position.

References

Anderson, S. B., Ball, S., Murphy, R. T., et al., 1976, *Encyclopedia of Educational Evaluation,* San Francisco/Washington/London, Jossey-Bass Publishers.

Broadfoot, P., 1979, *Assessment, Schools and Society,* Methuen and Co., London/New York.

Frith, D. S., Macintosh, H. G., 1984, *A Teacher's Guide to Assessment,* Cheltenham, Stanley Thornes.

Morgenthau, H., 1945, *The Evils of Politics and the Ethics of Evil,* in: Ethics (October 1945), pp. 14-17.

Mehrens, W. A., Lehmann, I. J., 1978, *Measurement and Evaluation in Education and Psychology,* New York, etc., Holt, Rinehart and Winston.

Newmann, F., 1950, *Approaches to the Study of Political Power,* in: Political Science Quartely, Vol. LXV, No. 2 June 1950, pp. 161-171.

12 Why Assess? Evaluating the Assessment of the English National Curriculum

C. G. Robinson

Introduction

In order to place the evaluation of the assessment aspects of National Curriculum of England and Wales into perspective I need to start by giving a brief summary of the National Curriculum as it is now and as it is at the moment intended to be in the future. The National Curriculum was set up under the 1988 Education Reform Act and the Act lays down that the Secretary of State for Education is required to introduce a National Curriculum as soon as practicable. This, the government is doing by introducing gradually, though far too fast for many observers and participants, different aspects of the National Curriculum.

Compulsory schooling in Britain spans the ages 5 to 16 which correspond to school years 1 to 11. For the purposes of the National Curriculum the years of compulsory schooling are divided into four key stages:

- Key Stage 1 spanning years 1 and 2;
- Key Stage 2 spanning years 3, 4, 5 and 6;
- Key Stage 3 spanning years 7, 8 and 9;
- Key Stage 4 spanning years 10 and 11.

Key Stage 4, therefore, spans the years which have been in the past – and indeed it is anticipated will be in the future – devoted to the General Certificate of Secondary Education or GCSE.

When the National Curriculum is fully implemented it will prescribe Programmes of Study and Attainment Targets for a range of 'foundation' subjects which will be compulsory for all students in the first three Key Stages and for the majority in Key Stage 4. So far however, only a small number of what are called 'core' subjects, English, mathematics, and science have been introduced with the full detail originally envisaged. A further area, technology, which in its new form will be new to the vast majority of schools, is due to be introduced at the beginning of the next academic year, and history and geography in the following September. Music, art, and physical education complete the foundation list.

The coincidence of Key Stage 4 and GCSE, and the earlier proposal from the Secretary of State that the main method of assessment at Key Stage 4 should be the GCSE, has led to some concern that the subjects of the National Curriculum will take over the whole of the available curriculum time and leave no room for other subjects which are currently taught and are chosen by many students in their final years of schooling. As a general rule, it is reckoned that a GCSE takes approximately 10% of the available teaching and learning time, and it doesn't take a mathematical genius to work out that if you have ten subjects required by law, there is little time for

anything else. The Secretary of State, therefore, has proposed that there should be some relaxation of the requirements of the National Curriculum in Key Stage 4, either by allowing some pupils who have attained high levels of attainment earlier to drop some subjects, or by providing a less demanding course for those pupils who do not wish to attempt a full GCSE. In the main core subjects – English, mathematics, and science – GCSE will be the basis of the assessment of the National Curriculum for all.

At the moment, schools are required to teach all the foundation subjects 'for a reasonable time' in the Key Stages 1 to 3. The requirement to teach for a reasonable time in Key Stage 4 applies only to the core subjects at present; for the remainder this requirement has been delayed until 1993 when there will be more details as to the needs of the syllabuses. From Key Stage 3 there is a requirement that all pupils should learn a modern foreign language but this does not form part of the requirement for the earlier key stages – perhaps a surprising decision in view of the rapid approach of 1992 with its stronger European links.

So the National Curriculum itself is being introduced gradually over a number of years and it will not be until 1994 or 1995 that all required subjects will be taught according to Programmes of Study as set down by government across all of the key stages; and it will not be until the turn of the century that the children who embarked upon the National Curriculum at age 5 in 1989 finally emerge from it at the age 16 having passed through all of its stages.

So what about the assessment?

The requirement to assess each child at the end of every key stage is laid down in the 1988 Act and cannot therefore be changed other than by modifying that Act. But the precise form of that assessment is to be defined in separate orders. And as yet we have no orders.

The government set up a group called the Task Group on Assessment and Testing (TGAT) in 1987-88 to advise on the most appropriate model for the assessment of the proposed National Curriculum. That group – chaired by Professor Paul Black – proposed a combination of teacher assessment and assessment through what are known as Standard Assessment Tasks. And so for the last two years, three different agencies have been working on the development of these Standard Assessment Tasks for Key Stage 1 and covering the requirements of English, mathematics, and science but approaching them in a cross-curricular setting. The current situation is that a sample of approximately 2% of the seven year olds in England and Wales has recently been using the Standard Assessment Tasks (or SATs) and over the next few weeks and months the lessons learnt from that exercise will be analysed and a report prepared for the School Examinations and Assessment Council when it meets in October 1990. At that point final arrangements will be made for the assessment of the entire 7 year old cohort 1991, and for future years.

As it stands at the moment, there are two aspects to the assessment: Teacher Assessment (TA) and SATs. Teacher Assessment is a review by the

teacher of the level of achievement of each child which is carried out on the basis of the teacher's own records and is completed during the spring term of the final year of the Key Stage. This is followed by SATs which are conducted during the first half of the last term of the academic year i.e in our system the summer term, in other words during the latter part of April and most of May.

So what is a SAT?

Essentially a SAT is a set of teaching materials designed to ensure that, in the course of doing the activities, the pupils will provide the teacher with evidence of their current attainment on a particular aspect or aspects of the National Curriculum. The focus is one or more Statements of Attainment (SoA), against which success is to be judged. To the extent that the SoA are criteria, the system of assessment is criterion-referenced. But it is not the success or failure on the SoA that is reported, as might be expected in a true criterion-referenced system. The SoA are grouped into sequential 'levels', and it is on the basis of the child's pattern of successes across all these SoA that the child is judged to have achieved that level, or not.

At the moment there are three subjects being assessed – English, mathematics, and science. These are the core subjects of the National Curriculum and the Government has recently announced that it does not intend the remaining foundation subjects – technology, history, geography, music, art, and physical education – to be subject to such a heavy assessment regime. Precisely how those subjects are to be assessed is yet to be determined.

Evaluating the National Curriculum Assessment

The evaluation must address the following questions.

- Is the assessment practicable?
- Is the assessment a valid reflection of the curriculum?
- Is the assessment reliable?
- Are the uses made of the assessment beneficial?
- Are there any damaging effects of the assessment?

It is, of course, too early to offer results but I can indicate where we are looking and what evidence we shall be seeking in an attempt to answer each question, as well as some early thoughts on the direction we are taking. These are related to the assessment itself but there are other questions more broadly based, and, therefore, of interest to the National Curriculum Council and the Curriculum Council for Wales as much as they are to SEAC.

- Are the statements of attainment appropriate summaries of learning objectives?
- Do the levels represent a realistic progression for the majority of children?

- Do the attainment targets and the profile components offer an appropriate reflection of the subject as a whole?
- Does the National Curriculum and its associated assessment improve the education of the children involved?

While these essential questions must, of course, be tackled, I shall concentrate in this chapter on questions relating directly to assessment.

Is the assessment practicable?

In short, the answer at the moment has to be 'No'. But we are still in the early stages and have time to consider. The agencies are analysing their data to try to identify the best way of reducing the burden of assessment.

I don't think there is a real problem with the Teacher Assessment – that is the assessment that takes place before the Standard Assessment Tasks are used. What may be a problem is misinterpretation of what is meant by the requirement to 'assess' and particularly of what is meant by assessing 'at the end of the Key Stage'.

Teachers need to assess their pupils to find out what they are learning. Because there is now a National Curriculum, which lays down what pupils are supposed to be learning, it is sensible to expect teachers to assess their pupils on the targets of the National Curriculum. The problem of over-burdening teachers with assessment arises when that assessment is constrained, that is when teachers are required to make their initial observations – on which the assessments are based – in too short a period.

What is sought is a system in which, during the course of the instruction over a period of two or three years, teachers make systematic observations of pupils and record their achievements. Then, at the end of the Key Stage (or indeed at the end of any reporting period) the teacher will need to decide how those observations relate to the requirements of the NC Attainment Targets. At this point the teacher will need to check how far observations lead to a judgement that the SoA for particular Attainment Targets have been achieved. By the application of some aggregation rules comparing achievements in adjacent levels, the teacher can come to a further judgement as to the level that best reflects the achievement of the pupil. That is not a burden – it is what teachers have always done – but it is systematized by reference to SoAs. It is only the last part of this process that needs to be done 'at or near the end of the Key Stage' – a review of the child's overall achievements in the light of observations made earlier and I believe that, provided the criteria for that final review are clear and realistic, teachers will find the time spent useful and not a burden.

Any formalised system of externally devised assessment will have some disruptive impact upon the normal classroom situation. What we have to do is ensure that the benefits outweigh any losses. We cannot yet do that. The trials that have taken place over the last two years have produced:

- teaching and assessment materials that are widely applauded;
- innovative approaches to assessment;
- methods of data handling.

But there is still an almost universally held view that the amount of time that teachers are having to give to the SAT assessment is too great, possibly twice as much as would be reasonable. (The burden on individual pupils does not seem to be a major issue; children find the materials enjoyable and not threatening.) If we are to reduce the burden on teachers, there must be a revision of the programme. We are still awaiting the results of the pilot assessments and the reports from the agencies but the choices appear to be:

- a reduction in the number of Attainment Targets addressed in the SAT, or
- a reduction in the number of Statements of Attainment addressed in the SAT, or
- a reduction in the number of pupils assessed in the SAT, or
- a reduction in the number of contexts used, or
- a combination of some or all of the above.

Against all of that we will need to balance the requirement that the assessment needs to be reliable and valid and that it has a beneficial effect upon the individuals who are assessed and on the system as a whole. That is going to be a difficult order and we must be ready to check how far we are succeeding, and be ready to modify the process to achieve the necessary improvements.

There are problems in all of these. Reducing the number of Attainment Targets addressed by the SAT will reduce the potential of the SAT as a check on the Teacher Assessment. It is assumed, at the moment, that all pupils will be assessed by SATS and any suggestion for the sampling of pupils would need to be backed by evidence that the Teacher Assessment was sufficient and comparable for the remainder. Finally the number of contexts in which a child is assessed may be reduced but again it would be necessary to show that the context(s) selected gave sufficient evidence for a generalisation of attainment.

I have spent a long time on the problem of manageability because that is what is currently highest in our priorities and because we have some feedback already. Such is not the case on many of the other questions and I can only indicate what we – by which I mean either the Evaluation and Monitoring Unit at SEAC, the SAT development agencies or the evaluation team at Leeds University – expect to address and the evidence we are seeking.

Is the assessment a valid reflection of the curriculum?

If it is not, the results cannot provide useful information for the teacher, pupil and parent which is the essence of a formative system. To some extent, the validity of the assessment is determined by the validity of the SoA laid down in the Order. However good the assessment of the SoA itself may be, it is not valid if the SoA is not a true reflection of the programme of study (which it should be, having been devised at the same time by the same people) and of the teacher's interpretation of that programme of study.

This is perhaps the main question as far as validity is concerned. Is it possible to devise valid assessment against nationally prescribed SoAs and Programmes of Study which inevitably leave a great deal of flexibility in interpretation to individual teachers?

The fact of the matter is that the whole of the assessment system relies heavily – currently too heavily! – on the teacher. The child's teacher carries out all the assessment, both TA and SAT, and therefore, unless the mode of assessment in the SAT is severely constrained, the teacher will inevitably use the same interpretations in the assessment as in the course. In relating the assessments to the individual pupil or teaching group, it is likely that there is a fairly high validity – an essential feature if the assessment is to be useful in feeding back formative information.

Is the assessment reliable?

Unreliable assessment is useless for any purpose whatsoever. But reliability has a number of aspects that are not always equally attainable. If two children's attainments are the same, the assessment of those two children on those attainments should be the same. This can be applied to a single teacher or to many. We may compare the individual teacher's judgement of two children with other teachers' judgements of the same children and would expect a high degree of agreement. We may compare the individual teacher's judgement of a child's performance on one occasion with that same teacher's judgement of the same performance on another occasion – again we expect a high degree of agreement. But there is a major problem of reliability built into the degree of interpretation that I have already argued adds to the validity of the assessment. Teachers may be shown, individually, to be highly consistent in their interpretation of the SoA and their judgement of the children in their class. They may at the same time be shown to be interpreting the SoA in different ways when compared with other teachers. This leads to a question of the priority given to the purpose of the assessment.

If the major function of the assessment is to inform the individual teacher of the progress of the children in that group then it is essential to have consistency within the group but less important to establish common standards across teachers. The more there is a requirement that the results of different groups are comparable, the greater is the need to ensure that teacher's interpretations of the SoA are the same.

Are the uses made of the assessment beneficial?

This is a longer term problem which will necessitate cooperation between ourselves and other bodies such as NCC and HMI. The hope, of course, is that the backwash effect of the assessment programme will help teachers to improve the learning experience of the children in their care. But we must check. The major problem that I foresee in this area is the belief that the

assessment has some predictive validity. This is unlikely to be the case and, to a large extent it would be dangerous if it were as there would be a temptation to regard the child as being 'a level 1 child' or 'a level 2 child' and so on. It is the child's performance on a particular AT that may be described as level 1 or level 2, not the child. And that performance is an indication – we hope a good one – of what the child can do at the time of the assessment, not an indication of 'potential'.

Of course we must expect teachers to make use of the information they have obtained from the assessment as the starting point for future work. That is wholly appropriate provided it is recognised that progress from that point will differ for individuals. No doubt the results will be used for other purposes – monitoring standards within schools, Local Education Authorities and nationally, and we will need to keep all those uses under review to check for their effects at each level. Already Local Education Authorities are working on currently available data, information on the pupils' attainment on entering a school compared with their attainments on leaving, to discuss what additional information on the individual children and on the school as a whole is necessary to try to isolate the achievements of the school itself. The work on Multi-level Models being carried out by Professor Harvey Goldstein at the London Institute of Education is a useful base for these investigations.

Are there any damaging effects of the assessment?

Assessment takes time. It is proposed that the materials provide not only opportunities for assessment but appropriate learning experiences. But that is difficult to achieve. When does teaching stop and assessment begin? What is the difference between explaining what the activity is asking the child to do and telling the child what to do? One is acceptable, the other is not. At least in the early stages, therefore, the main dangers lie in the effect the assessment role has on the relationship between teacher and pupil. It would be terrible if pupils find that their normally friendly teacher, who is always willing to help if you don't know how to spell a word or can't work out a mathematics problem, suddenly turns into an unfriendly, time-obsessed beast who refuses to give any word of encouragement lest the assessment be declared void!

Select bibliography

Department of Education and Science and the Welsh Office, 1989, *Science in the National Curriculum*, London, HMSO.

Department of Education and Science and the Welsh Office, 1989, *Mathematics in the National Curriculum*, London.

Department of Education and Science and the Welsh Office, 1989, *English in the National Curriculum*, London, HMSO.

Department of Education and Science and the Welsh Office, 1990, *Technology in the National Curriculum*, London, HMSO.

Department of Education and Science and the Welsh Office, 1989, *Task Group on Assessment and Testing: A Report*, London.

Department of Education and Science and the Welsh Office, 1989, *National Curriculum – From Policy to Practice*, London.

Education Reform Act 1988, 1989, London, HMSO.

13 Subject Pair Comparison Studies in an International Context

J. J. Thompson, M. C. Hayden, I. J. Kerr

Background: the International Baccalaureate

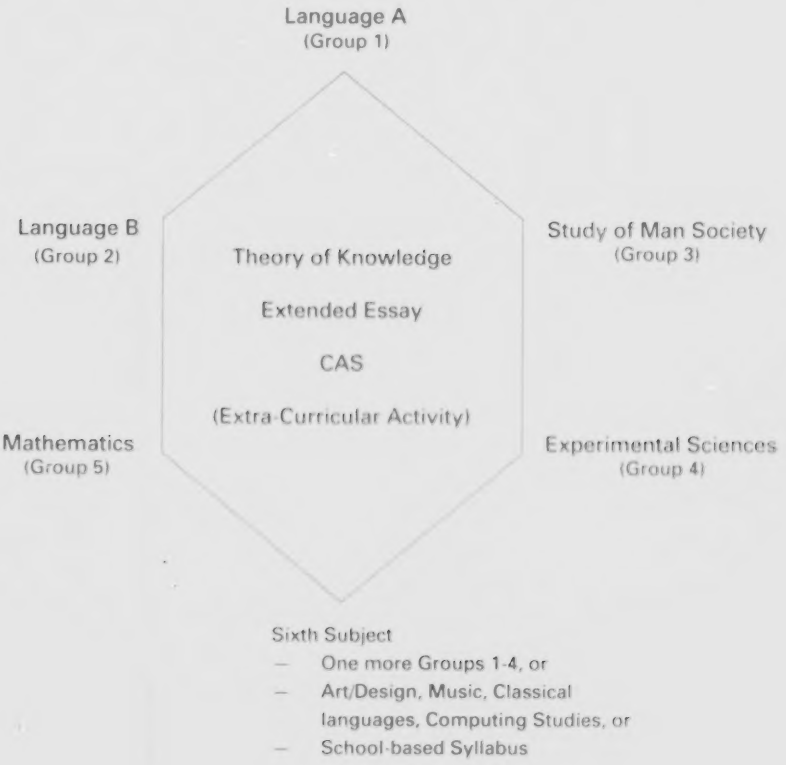
The International Baccalaureate is an educational programme developed in the 1960's as a means of catering for the children of employees of multi-national organisations, of diplomats, and of others who were internationally mobile and whose lifestyle often led to their offspring's formal education taking place in more than one national system. Problems of transferability and acceptability frequently resulted, and the concept of the International Baccalaureate was developed by a group of educators, largely based in Switzeriand, the U.S.A and the U.K, as a system which would have international currency.

The International Baccalaureate (IB) was designed to be followed by students in the 16-19 age range and to be the best form of compromise between the perceived extremes of specialisation and over-generalisation exhibited in the range of national systems. A structure of five groups of disciplines, from each of which one subject must be chosen, provides breadth of experience, while the choice of a sixth subject drawn either from a further list of options or from one of the groups already sampled, allows for specialisation. Specialisation is also represented in the offering of either 3 (or 4) subjects at Higher Level and the remaining 3 (or 2) at Subsidiary Level, as well as by the compulsory Extended Essay comprising some 4000 words, which encourages the student to undertake independent research in an area of particular interest. Coherence is provided through the compulsory Theory of Knowledge course, which encourages students to reflect on the different approaches to learning encountered in the different subjects studied, and the 'CAS' (Creativity, Action, Service) component required of each student involves activity in non-academic areas. These features of the International Baccalaureate 'package' have been represented pictorially by a hexagon (see figure 1 on the next page).

All academic subjects studied in the IB are graded on a scale from 1 (minimum) to 7 (maximum), with a variety of techniques being purposefully employed in the assessment process. Theory of Knowledge (TOK) and the Extended Essay are both graded on a system of bonus/penalty points, while CAS is not formally assessed in the same way (although each student's participation is monitored by the school). The IB Diploma is awarded to students whose total points score from six subjects, TOK and Extended Essay, reaches a minimum of 24 (with certain restrictions). A Certificate listing achievement is awarded to those who fail to reach this target: candidates may also register for only parts of the programme, leading to the award of a Certificate.

The IB has grown to the point where it is currently offered by almost 400 state and international schools in 56 countries and the Diploma is

Figure 1 The International Baccalaureate programme structure



accepted as an entrance qualification for universities and higher education institutions in around 60 countries of the world. Examinations are offered in English, French and Spanish at sessions in May (largely Northern hemisphere schools) and November (Southern hemisphere schools); approximately 13,500 students were registered to take examinations in the May 1990 examination session, comprising in the region of 42,000 subject entries.

Background to study

Built into the IB system of assessment is an unwritten assumption that there is comparability between a particular grade in one subject and the same grade awarded in another. Whilst no explicit claims to this effect are made by the International Baccalaureate Organisation (IBO), the fact that the same scale is used in grading all subjects and that these grades are then added together suggests that there is – or should be – a certain equivalence between, for instance, a grade 5 awarded for History at Subsidiary Level (SL) and a grade 5 awarded for Physics at Higher Level (HL). Although

consideration has previously been given to this issue within the IB system, no formal study of comparability had ever been commissioned prior to 1988, when a joint investigation was set up between the IBO and the School of Education at the University of Bath. The principal investigators were from the Research & Development section of IBO (Hayden), the Yokohama International School (Kerr, sabbatical leave) and the School of Education, University of Bath (Thompson).

Aim and scope of the study

The aim of the study was to investigate the extent to which established methods of evaluating comparability between subjects, as practised in English, Scottish, and Welsh Examination Boards, were effective when employed in the more heterogeneous population of candidates entering for the IB examinations. The scope of the study was the investigation of three major approaches to the consideration of subject comparability:

- Subject Pair Comparisons, considering differences in mean grades of pairs of subjects;
- the Consensus method, considering differences between means of particular subjects and a 'basket' of other subjects;
- correlation between results of different pairs of subjects.

In addition, it was intended to consider whether results when applied to IB data were similar to those of studies previously undertaken by other organisations (cf. the work of Nuttall, D.L., Backhouse, J.K., and Willmott, A. S., 1974, and of Forrest, G. M., and Vickerman, C., 1982). Whether the exercise would be sufficiently meaningful to justify the incorporation of such data as part of the regular analysis of examination results by the IBO, would also be considered.

Investigation 1: subject pairs method

Rationale

The basic argument supporting the comparison of subject grades is put by Nuttall, Backhouse and Willmott as follows: "We do not expect an individual candidate to achieve the same grade in every subject that he takes. However, we can see no logical reason why, if a large group of candidates representative of the population took, for example, both English and mathematics, their average grades should not be the same. (...) Although some individuals in the group will be better at English than they are at mathematics, there is no reason to suppose that these individuals will not be balanced by another group who are better at mathematics than they are at English. We therefore argue that their mean grades should be the same."

Forrest and Vickerman, in a study based on data from the Joint Matriculation Board (JMB) in Manchester (UK), modified this argument

significantly as follows: "...there is no logical reason why, if a large group of candidates representative of the population took", for example, two related subjects say Physics and Mathematics, "their average grades should not be the same."

Assumptions made of similar grade distributions, of grading systems not being strictly criterion-referenced, and of – on average – equally good teaching in the studies of Nuttall, Backhouse and Willmott and of Forrest and Vickerman, were considered also to be reasonable in an IB context. Furthermore, an assumption regarding the equality of motivation was accepted, not only for related subjects, but also for all other subjects attempted, on the basis that every point gained contributes equally to the candidate's total score.

Selection of data

Data were available from both the May 1988 and May 1989 examination sessions. With reference to the Forrest and Vickerman study, which, to ensure reliability of data, advocated a sample of at least 50 participants in the subject pair under consideration, a decision was made about what data should be used in the study. In order to allow for possible comparisons by gender, only subject pairs having at least 100 candidate entries would be considered. For similar reasons, only subject entries in Groups 3 to 6 with English as a working language were selected, since French and Spanish entries are relatively small in comparison. (Languages A and B are examined in the language in question and do not therefore have a 'working language' in the same sense.) Although, to our knowledge, previous studies had compared only subjects at the same level ('O' Level with 'O' Level for instance), for this study it was decided also to include 'cross level' pairs (i.e. HL with SL). While we had some reservations about this modification, it was felt that the fact that scores at each level contribute equally towards the total score, implies some parity between the two levels and was a justification for considering these data.

Appendix 1 shows the 50 subject pair combinations selected with a minimum of 100 candidate entries and English as the working language (for subjects other than Languages A and B) from the May 1988 examination session. Candidate entry figures are given in each case by total as well as broken down by gender.

For May 1989 examination session data, 54 subject pairs satisfied the same criteria (more than 100 candidate entries and English only as the working language in Groups 3 to 6): appendix 2 shows the breakdown of candidate entry figures for these pairs from May 1989, again by total as well as by gender.

Approach

For each pair of subjects selected, t-tests were carried out to determine whether or not the group of candidates selecting that pair were representative in each case of the total candidate entry for each subject. This

was felt to be important if any general conclusions about subjects were to be drawn from the pairs under investigation.

Consideration was also given to the issue of gender. Would it be meaningful to compare mean grades of a pair of subjects, without investigating whether in fact the difference between mean grades of boys and girls in one of the subjects might be greater than the difference between the mean grades of that and the other subject, both taken for the population as a whole? We felt that this aspect of the data should also be considered.

For both subjects in each pair, therefore, and in each case treating boys and girls as separate subgroups, mean grades were calculated for the subgroup of candidates taking the subject as part of the pair, and for the total group of candidates taking the subject. The following example for the subject pair English A SL (ES) and History SL (HS) illustrates the method.

	Boys				Girls		
	ES	HS	DP		ES	HS	DP
MP	4.82 (61)	4.39 (61)	0.43	MP	4.76 (50)	3.88 (50)	0.88
MA	4.58 (630)	4.76 (380)	-0.16	MA	4.69 (623)	4.01 (420)	0.68
DM	0.24	-0.37		DM	0.07	-0.13	

MP: mean grade for candidates taking subject in the pair
MA: mean grade for all candidates taking the subject
DM: difference between these means (DM = MP-MA)
DP: difference between subject means (e.g. in above case, DP = ES-HS)

Numbers in parentheses are the numbers of candidate entries in each case.

Sample population: May 1988 data

Using this method, a total of 100 subject pair grids (50 each for boys and girls) ware drawn up, 7 of which could not be used because the size of the subgroup was less than 50. The remaining 93 subject pairs fell into three categories as follows:

- Those where the subject pair mean grade difference (DP) was high or low, but for either or both subjects the subgroup which chose the pair was not representative of the total population (i.e. DM high in one or both subjects); those falling into this category were eliminated from further investigation.
- Those where the subject pair mean grade difference (DP) was low (determined by t-tests) and the subgroup was representative of the total population for both subjects (DM low in both cases).
- Those where the subject pair mean grade difference (DP) was high (determined by t-tests) and the subgroup was representative of the total population for both subjects (DM low in both cases).

Elimination of the first category left 28 subject pairs with DM low for boys or for girls, or for both: 41 grids remained of the original 93. It had already been noted that Language B results in general had particularly high mean grades (for example, for boys and girls together: French B SL : 5.01, French B HL 5.26, Spanish B SL : 5.06). Language B subjects have always posed particular problems regarding standards since, while they are designed to be a foreign language learning experience for students, it is known that there is abuse of the system in some cases by students registering for Language B examinations in which they are native speakers. Since this unintended use of the system could result in inflated mean grades, it was decided to omit at this stage all pairs containing a Language B at either level. Of the original pairs selected, the final list of 19 pairs which was felt to be meaningful to consider further, is included in appendix 3.

Sample population: May 1989 data

The same exercise was repeated for May 1989 data. Having eliminated those pairs with fewer than 50 candidate entries, 96 pairs fell into the three categories already identified. Pairs from the first category were again eliminated, and t-tests were run to determine whether DP and DM were 'high' or 'low' in each case. Language B subjects again produced particularly high mean grades (French B SL: 4.98, French B HL: 5.38, Spanish B SL: 5.05) and all Language B subjects were again eliminated at this stage. For the May 1989 data the final list which it was felt was meaningful to consider, consisted of 15 pairs and is also included in appendix 3.

Results

Of the 19 subject pairs considered from the May 1988 data, two (English A SL with Economics HL and History (Europe) HL with Mathematics SL) showed close agreement (i.e. their mean grade differences were not significant at the 5% level) for both gender groups. Four pairs:

- English A SL with History (Europe) HL,
- English A HL with Economics HL,
- English A HL with Physics HL, and
- English A HL with Mathematics SL

showed a significant mean grade difference for one gender subgroup but not the other.

Four other pairs:

- English A HL with Biology HL,
- English A HL with Chemistry HL,
- History (Europe) HL with Biology SL, and
- Biology HL with Mathematics SL

showed a significant difference in mean grades for both gender subgroups.

The remaining 9 pairs provide information for one gender subgroup only since the other subgroup for the pair in question had been rejected on the basis that it was not representative of the total population; 6 of these 9 showed significant differences.

Two points in particular are worth noting:

- Of the 13 subject pairs in which English A appears, at either level, 11 show a significantly large mean grade difference for either the boy or girl subgroups, or both.
- Where there is a significant difference in mean grades between subjects, this occurs on 13 occasions for the girl subgroup, but on only 5 occasions for the boy subgroup. (Of the 4 pairs showing a difference for one subgroup but not the other, it was in each case the girl subgroup which showed the difference).

It may be that these points suggest a particular discrepancy in standards between English A and other subjects; it may also be that further work should be undertaken to investigate whether certain examinations show gender bias.

Appendix 3 incorporates the lists from both May 1988 and May 1989. It will be noted that the same list of subjects from May 1988 was not duplicated in May 1989. Four new pairs of subjects:

- English A SL with Biology SL,
- History (Europe) HL with Chemistry HL,
- Economics HL with Biology SL, and
- Chemistry HL with Mathematics SL

were added to the list, while 8 pairs which had been shortlisted in 1988 as having subgroups representative of the total population were no longer included.

Again, two points are worth noting from the May 1989 data:

- English A, at either level, appears in 8 subject pair combinations; of these, 5 show a significantly large mean difference for either the boy or girl subgroups, or both.
- Where there is a significant difference in mean grades between subjects, this occurs on 9 occasions for the girl subgroup and on only 3 for the boy subgroup.

While the outcomes for the May 1988 and May 1989 samples are different in specific respects, the same patterns seem to have emerged. English A (at either level) shows a larger number of discrepancies with other subjects than any other subject, and the girl subgroups show a larger number of significant discrepancies overall than those of the boys. The emergence of this apparent pattern reinforces the notion that further work would be justified in investigating these two areas in particular.

Shortcomings of the approach

In drawing conclusions from the data produced, it should be taken into account, that some subject pairs have been rejected as providing meaningful data by dint of the fact that their subgroups were not representative of their total population. Those which were not rejected at this stage, which were not only representative but which also had large enough samples to be meaningful, appeared in different numbers of combinations.

Care should also be taken in, for instance, drawing conclusions about the standards of English A from the number of significant differences appearing with other subjects, since in the 19 subject pairs eventually considered, English A at one level or the other appeared 13 times, compared with, for instance, the next highest, History, which appeared only 7 times. If more of the particular pair combinations for which data were available had been representative of their parent populations, interesting results in other subjects may also have emerged.

Investigation 2: consensus method

Approach

The consensus method was described by Nuttall, Backhouse and Willmott in 1974. It attempts to make a judgement about the relative 'leniency' or 'severity' of a particular subject by comparing its mean with the mean grades of a 'basket' of other subjects, taken one at a time. If the mean of the means is lower than the mean of the mean grades of the 'basket', the subject in question is judged to be harsh, and vice versa. The method was applied to IB data from both May 1988 and May 1989 in order to attempt to calculate the 'leniency' or 'severity' of English A SL and Mathematics SL, in each case for the population as a whole. For both these subjects, subject pair groupings greater than 100 could be identified with ten other subjects.

Results

Results arising from application of the consensus method to May 1988 data suggest that Mathematics SL is 0.04 of a grade 'lenient' (4.58-4.54) with respect to the consensus, and that English A SL is 0.26 (4.79-4.53) of a grade 'lenient' with respect to a slightly different consensus.

The same exercise using May 1989 data for the same combinations of subjects produces similar results although, interestingly, this time the results suggest that Mathematics SL is 0.03 of a grade 'severe' (4.54-4.57) while English A SL appears only 0.02 of a grade (4.77-4.75) 'lenient'. The only patterns which might perhaps be noted for the two years are the similarities in mean grades for Mathematics SL and its associated consensus from 1988 to 1989, and in mean grades for English A SL: the mean grade of the consensus for English A SL meanwhile shows an increase from 4.53 in 1988 to 4,75 in 1989.

If the above data were to be taken at face value, the difference in mean grade between English A SL and the consensus might perhaps give rise to concern. We did however have some reservations about this method.

Shortcomings of the approach

This measure of 'leniency' or 'severity' of one subject compared with a group of others suffers from the fact that it is unweighed, in the sense that the means of a pair combination, which is a very popular choice, contribute equally to the consensus with the means of a pair combination which has been selected by only a few candidates. Clearly too, the 'measure' depends on which subjects are selected to contribute to the 'basket', and is therefore to some extent arbitrary. In considering the possibility that gender differences, hidden in any comparisons of overall populations, might be worth investigating, further cause for concern was uncovered. English A HL (EH) in May 1988 is chosen as a good example here of the source of that concern:

	Boys and Girls		Boys only		Girls only	
	EH	Others	EH	Others	EH	Others
Mean grade	4.89	4.40	4.47	4.60	5.02	4.23
Difference	0.49		0.14		0.79	

Each of the gender subgroups had first been shown to be representative of the total gender population taking the subject in question. If the total population is then considered, English A HL appears almost half a grade 'lenient' compared with a consensus; a consideration of the results broken down by gender subgroup however indicates that for the boys alone English A HL is only 0.14 of a grade 'lenient' while for the girls alone English A HL appears to be 0.79 of a grade 'lenient' when compared with the same consensus. The possibility of large discrepancies between male and female subgroups being overlooked in any use of the consensus method for total populations gives cause for concern regarding the reliability of this method unless such factors are taken into account.

Investigation 3: correlation coefficient method

Approach

In this investigation, Pearson Product-Moment correlation coefficients were calculated from both May 1988 and May 1989 data for all pairs of subjects having at least 100 candidate entries in common; no account was taken of gender difference.

Results of the May 1988 data

Fifty pairs of subjects in the May 1988 examination session had common candidate entries of at least 100: coefficients for these 50 pairs ranged from 0.30 for Spanish B SL with English A HL (considered by interpretation of most of the literature, including the study of Forrest and Vickerman, to be weak) to 0.78 for Biology HL and Chemistry HL (considered strong). Perhaps not surprisingly, the correlation coefficients derived show similarity to those of the Forrest and Vickerman JMB studies, with highest correlations occurring between allied subjects (such as two sciences, or a science with mathematics). The lowest correlation between allied subjects was found between Biology SL and Mathematics HL with a coefficient of 0.51. Coefficients involving English A at either level were generally higher when compared with 'arts' subjects such as History than with a science or mathematics, although English A SL compared with Mathematics SL and Biology SL produced coefficients of 0.52 and 0.53 respectively, only slightly lower than the correlation with History at either level. Interestingly, an association pointed out by Forrest and Vickerman, of Economics with most other subjects, was also highlighted in the IB study which produced coefficients greater than 0.5 for all pairings with Economics HL except the three Language B subjects. Only one pair with an entry greater than 100, French B SL with Biology SL, produced a coefficient found not to be significant: the correlation coefficient here was 0.14. In general, combinations involving a Language B produced low coefficients.

Results of the May 1989 data

In applying the same method to May 1989 results, a similar pattern of correlation coefficients emerged. In particular, the six subject combinations highlighted for comparison with results from the JMB produced similar coefficients with May 1989 data as for May 1988 data and, indeed, with JMB 1980 data: a summary of the results for these six subject pairs appears as appendix 4.

The similarity between IB results for both 1988 and 1989, and those of Forrest and Vickerman (albeit in a limited number of subjects and comparing IB data from 1988 and 1989 with JMB data from 1980) was interesting (see appendix 4), and reassuring in the sense that it might perhaps be interpreted to mean that there is something common to these particular subject areas which overrides other issues more specific to particular Examining Boards.

Shortcomings of the approach

Although this was an exercise of interest in terms of comparing results from year to year and highlighting similarities with JMB results, it is difficult to know to what use the results can be put. High levels of correlation do not of themselves demonstrate cause and effect, nor do they give any absolute indication of standard. The results of this exercise are unlikely to contribute

substantially to any investigation of standards within the IB, in the future.

Further work

It is intended that work should proceed in the following principal directions:

- an extension of the current analysis to a wider range of IB subjects;
- an extension of the analytical methods used, to include Analysis of Variance (ANOVA);
- further investigation of gender differences in subject comparability (with particular reference to work published by G.M. Forrest in 1988);
- an extension of the method to investigate further work undertaken recently by other organisations on grade by grade comparisons of results;
- an investigation of the use to which data from such studies can be put in the setting of questions and the composition of question papers;
- an investigation of the ways in which the results of the application of such methods may contribute to grade awarding procedures throughout the IB;
- an enquiry into the extent to which such data can be used in the orientation and development of examiners within IB through training sessions and materials.

References

Forrest, G. M., 1988, *Subject Comparability and Gender Difference in GCE Advanced Level Examinations*, University of Keele.

Forrest, G. M., Vickerman, C., 1982, *Standards in GCE: Subject Pairs Comparisons 1972-80.*

Nuttall, D. L., Backhouse, J. K., Willmott, A. S., 1974, *Comparability of Standards between Subjects.*

Appendix 1

Subject Pair Combinations with 100+ entries, total entry, boys entry, and girls entry respectively; May 1988 data (English working language)

	ES	EH	FS	FH	HS	HH	EC	BS	BH	CH	PH	MS	MH	SS
ES	*	*	181		111	149	177	139	221	276	262	209	333	
			91		61	74	113	53	102	170	191	113	211	
			87		50	73	61	83	116	105	69	95	121	
EH	*	*	659	174	195	647	200	340	538	365	211	846	260	242
			240	55	75	268	92	92	193	209	146	396	139	112
			415	118	118	376	107	248	341	156	63	445	121	130
FS			*	*		291		114	299	257	168	398	172	
						107			93	141	106	184	98	
						182		92	203	114	60	210	74	
FH			*	*				116						
								80						
HS					*	*						119	129	
												54	77	
												64	52	
HH					*	*	108	240	235	108		358		
							66	85	100	67		191		
								154	133			166		
EC							*	120				179	162	
												95	112	
								77				83		
BS								*	*			191	108	
												83		
												107	69	
BH								*	*	236		426	146	
										111		192	64	
										125		231	82	
CH										*	340	333	327	
											253	187	232	
											86	145	95	
PH											*	247	398	
												172	310	
												71	86	
MS												*	*	
MH												*	*	
SS														*

* Subject pair combination not allowed

ES:	English A SL	HH:	History HL (Europe)	PH:	Physics HL
EH:	English A HL	EC:	Economics HL	MS:	Mathematics SL
FS:	French B SL	BS:	Biology SL	MH:	Mathematics HL
FH:	French B HL	BH:	Biology HL	SS:	Spanish B SL
HS:	History SL	CH:	Chemistry HL		

Appendix 2

Subject Pair Combinations with 100+ entries, total entry, boys entry, and girls entry respectively; May 1989 data (English working language)

	ES	EH	FS	FH	HS	HH	EC	BS	BH	CH	PH	MS	MH	SS
ES	*	*	155	105	127	164	190	134	205	301	291	248	360	
			77	53	72	85	123	52	79	174	226	133	228	
			78	52	55	79	67	82	126	127	65	115	132	
EH	*	*	767	181	188	755	219	363	639	359	250	979	273	287
			318	42	71	310	98	99	238	190	171	454	150	110
			449	139	117	445	121	264	401	169	79	525	123	176
FS			*	*	106	373		151	346	254	204	457	170	
					54	154		44	115	144	141	214	100	
					52	219		107	231	110	63	243	70	
FH			*	*				119				104		
								25				45		
								94				59		
HS					*	*		102	108			115	162	
								40	39			49	90	
								62	69			66	72	
HH					*	*	103	258	264	149		432	132	
							62	100	105	81		207	82	
							41	158	159	68		225	50	
EC							*	169				212	169	
								77				123	109	
								92				89	60	
BS								*	*			210	101	
												93	51	
												117	50	
BH								*	*	224		513	178	127
										89		217	74	50
										135		296	104	77
CH										*	330	382	333	
											246	209	212	
											84	173	121	
PH											*	312	401	
												225	312	
												87	89	
MS												*	*	168
														74
														94
MH												*	*	
SS														*

* Subject pair combination not allowed

ES:	English A SL	HH:	History HL (Europe)	PH:	Physics HL
EH:	English A HL	EC:	Economics HL	MS:	Mathematics SL
FS:	French B SL	BS:	Biology SL	MH:	Mathematics HL
FH:	French B HL	BH:	Biology HL	SS:	Spanish B SL
HS:	History SL	CH:	Chemistry HL		

Appendix 3

Final shortlist of pairs considered, May 1988 and May 1989 data

Subject Pairs	ES		EH		HS		HH		EC		BH		CH
HS	88		88	89	–		–						
HH	88	89			–		–				88	89	89
EC	88	89	88	89					–				
BS		89	88				88	89		89	–		
BH			88	89			88	89			–		
CH	88		88					89				–	
PH	88		88										
MS	88	89	88	89	88		88		88	89	88	89	89

ES: English A SL
EH: English A HL
HS: History SL
HH: History HL (Europe)
EC: Economics HL
BS: Biology SL
BH: Biology HL
CH: Chemistry HL
PH: Physics HL
MS: Mathematics SL

– : Pair combination not allowed

Appendix 4

Correlation coefficients; comparison of JM8 1980 and IB May 1988/1989 coefficients for six Subject Pairs*

	Subject Pair 1980	JMB 1988	IB 1989	IB
JMB	English Lit. – Maths. Syll. A	0.35		
IB	English A HL – Maths HL		0.39	0.37
JMB	English Lit. – Chemistry Syll. B	0.39		
IB	English A HL – Chemistry HL		0.40	0.41
JMB	English Lit. – History Syll. A	0.56		
IB	English A HL – History (Eur.) HL		0.54	0.46
JMB	Physics – Chemistry Syll. B	0.75		
IB	Physics HL – Chemistry HL		0.76	0.76
JMB	Physics – Maths Syll. A	0.76		
IB	Physics HL – Maths HL		0.61	0.67
JMB	Chemistry B – Maths Syll. A	0.70		
IB	Chemistry HL – Maths HL		0.70	0.74

* Comparison restricted to six pairs published by Forrest and Vickerman in 1982.

14 The Use of the Item Response Theory in the Dutch National Assessment Programme

N. D. Verhelst, T. J. H. M. Eggen

Introduction

In 1987, under the authority of the Dutch Ministry of Education, the National Institute for Educational Measurement (CITO) launched the Dutch National Assessment Programme (DNAP) in elementary education. The purpose of the programme is to report periodically on the achievement of elementary school students in the Netherlands as a whole, as well as subpopulations such as groups based on ethnicity and gender, in order to provide the basis for a more informed discussion about the contents and the level of elementary education. The first phase of the programme concerned Mathematics in grade 3 (9-year-olds) and grade 6 (12-year-olds). First, the framework in which the DNAP is executed is presented. Then it is argued that partly as a result of these characteristics, psychometric models commonly employed in national assessment programmes are inadequate. Therefore, two new psychometric models were developed, i.e., a measurement model and a structural model. Although the models were developed as part of a national assessment programme, they can also make an important contribution to the solution of many other problems in educational measurement. Finally, a description of the computer programmes for the analyses is presented in section five.

Framework of the DNAP

The characteristics of the DNAP are related to legislative and/or budget constraints and design considerations. As a result of a legal obligation to publish part of the items administered in the first assessment, the tests used for the second assessment cannot be the same as those used for the first. Because of the fact that schools are not obliged to participate and because of budget limitations only a sample of elementary schools is taking part in the programme. Owing to budget constraints, the calibration of the measurement instruments and the actual assessment had to be conducted on one and the same sample of students.

The DNAP had to make a detailed discussion of the results possible. The subject area domain Mathematics was therefore divided into 40 homogeneous subdomains or tests. Employing homogeneous tests has the advantage of making the assumption of unidimensional tests more probable, and gives some guarantees against the instability of tests over time. There were 13 tests (275 items) for the grade three sample and 27 tests (475 items) for the grade six sample. The items were not of the multiple-choice format but were dichotomously-scored exercises to prevent students

from guessing the right answers to the items.

In order to report on the achievement of relatively small subpopulations (e.g., ethnic groups), a disproportionate stratified sample was drawn. Within the strata, all the students in a relevant grade of a random sample of schools were selected to reduce the costs of the collection of the data. The sample for the third grade was selected independently of the sample for the sixth grade. To control for cluster effects, the tests for each grade were divided into six groups, and every student was administered items from one group only. This way one item was administered to about four students from the same school.

Increasing the precision of the parameter estimates as well as using the testing time per student efficiently was accomplished by having the teachers divide the sampled students into a high ability group and a low ability group, and having the test constructors divide the items into easy items, difficult items, and items of medium difficulty. The low ability group was administered easy items and items of medium difficulty. The high ability group was administered items of medium difficulty and difficult items.

The design used for each subdomain was thus an incomplete design whereby the two subsamples were not statistically equivalent. In Rubin's terminology (Rubin, D. B., 1976), this is an incomplete design in which the missing observations are not 'missing at random'. It is this latter characteristic which complicates the analyses considerably.

Separability of the measurement problem and the structural problem

Findings from other national assessment programmes have shown that item response theory is the most suitable framework to obtain the information assessment projects have been designed for. Given the above, the two-parameter logistic model would have been an obvious choice for the measurement model. However, there are two objections to this choice. The first objection is that for the two-parameter model there are no sound statistical tests to detect item misfit. Therefore well-founded decisions for the removal of items from a test cannot be made. The second objection is closely related to how parameters are estimated in computer programmes for the two-parameter logistic model. BILOG is probably the most widely used programme for this model. BILOG employs the marginal maximum likelihood (MML) method, whereby the measurement model has to be complemented with a model for the distribution of the latent variable. The estimation problem can best be illustrated for the normal distribution. BILOG assumes that the sample is a random sample from one normal distribution. However, in our case the low and high ability groups are clearly not equivalent. Therefore, performing the calibration with BILOG forces us to make a specification error which of necessity produces inconsistent estimators of the item parameters. This means that BILOG is not suitable for incomplete designs in which the missing observations are not 'missing at random'.

The problem, however, is much more general. Assume that there are two background variables with I and J categories respectively, dividing the

total population into IxJ subpopulations. Also assume that the latent variable in each of the subpopulations has a normal distribution with common variance. To obtain consistent estimators of the item parameters with the MML method, the structural hypothesis on the distribution of the latent variable has to be absorbed into the model. The result is that the IxJ means, the common variance and the item parameters have to be estimated at the same time, which is complicated but technically possible. However, problems arise if we obtain a bad fit for the model. A bad fit could be caused by one or more items and/or by a misspecification of the structural model such as the presence of an interaction between the two background variables and the latent variable. No statistical methods are known which can distinguish well between these two sources of misfit.

Because of the two objections, only measurement models that allow for the separate investigation of the measurement model and the structural model can meet our purposes. As a matter of fact, a measurement model satisfying our demands is a model with which the item parameters can be estimated without making assumptions about the distribution of the latent variable and with a sufficient statistic for the person parameter. A suitable estimation method producing consistent estimators is the conditional maximum likelihood (CML) method. An important additional benefit of the CML method is that the sampling method does not affect the consistency of the estimators. Whereas the MML method requires difficult and often inaccurate corrections of the estimators as the result of using a disproportionate stratified sample, the CML method is sampling independent.

Measurement model: the generalized one-parameter logistic model

The only model, discussed in the literature, meeting the demands mentioned is the Rasch model. This model, however, assumes all items to have equal discrimination parameters, which is not very realistic. By conceiving the discrimination parameters in the two-parameter model as known constants, however, a generalized one-parameter model is obtained with the score $s = \sum_i a_i x_i$ as the sufficient statistic for the person parameter. In s, a_i is the discrimination index of item i, and $x_i \varepsilon\{0,1\}$ the response variable for item i. To avoid trivial estimation equations, we have to assure that the response pattern does not become a one-to-one function of the scores defined this way. This can be accompiished by restricting the discrimination indices to positive integers. Although the existence of unique CML-estimators for arbitrary discrimination indices is not always guaranteed, in practice this does not cause problems if the number of items is not too small (10 or more items) and the variance of the discrimination indices is not too large ($2 \leq a_i \leq 10$, say). Consequently, the generalized one-parameter model is a family of models, whereby every family is characterized by a vector of difficulty parameters and fixed discrimination indices. The union of these subfamilies is a model space which almost coincides with the two-parameter logistic model. Its great advantage is that every member of this model space is an exponential family model with

sufficient statistics for the person parameter so that unique CML estimates of the item parameters can be obtained for most practical situations. Thus far only two times out of more than a thousand analyses, has a model been specified without unique CML estimators.

From a statistical viewpoint, fixing the discrimination indices amounts to a specification of the null hypothesis which is stronger (i.e., more specific) than the null hypothesis in the two-parameter model. We therefore found it necessary to develop statistical tests which were sensitive to misspecifications of these indices. Statistical tests in which the test statistic is asymptotically chi-squared distributed with one degree of freedom were developed by Glas (Glas, C.A.W., 1989), and Verhelst and Glas (Verhelst, N.D., Glas, C.A.W.,1990). Their results specify under which conditions quadratic forms are asymptotically chi-squared distributed and at the same time permits the construction of quadratic forms targeted at testing against specific alternative hypotheses. The three tests (per item) developed are very sensitive to a misspecification of the discrimination index of the item, and immediately indicate the direction of the misspecification. If significant, discrimination indices are adapted and analyses repeated. In most cases satisfactory results were obtained after three or four analyses. A fourth test, also an asymptotically chi-squared distributed statistical test, was developed to provide information on the overall fit of the item in the model. Of the 40 tests that were calibrated this way, less than five percent of the items had to be removed because of misfit.

Structural model: the linear model applied to the latent variable

After the measurement model has been accepted, the item parameters are assumed to be known and the structural model can be investigated. An example of a structural model with two background variables is the following linear model:

$$\theta_{ij} = \mu + \alpha_i + \beta_j + \varepsilon \quad (i=1,\ldots,I;\ j=1,\ldots,J) \qquad (1)$$

where θij is the latent variable, μ is an additive constant, α_i is the effect of background variable 1 at level i, β_j is the effect of background variable 2 at level j, and ε is the error term with $\varepsilon \sim N(0,\sigma)$. An arbitrary technical restriction is that $\alpha_1 = \beta_1 = 0$.

The fundamental difference with the classical linear model is that the dependent variable is not observed. Replacing θij in equation (1) by an estimate of θij and applying the linear model would lead to inconsistencies in the parameter estimates (i.e., regression towards the mean). The likelihood function of a response pattern x_{ij} of a randomly selected student from the ij-th subpopulation is

$$L(\mu,\alpha_i,\beta_j,\sigma;\underline{x}_{ij}) = \int_{-\infty}^{\infty} h(\underline{x}_{ij}|\theta)\, g_{ij}(\theta)\, d\theta \qquad (2)$$

$$g_{ij}(\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp -[\theta - \alpha_i - \beta_j]^2/2\sigma$$

In equation (2) $h(x_{ij}|\theta)$ is the conditional probability density of response pattern $\underline{x}_{ij}$ under the generalized one-parameter logistic model. Maximization of the likelihood function for all response patterns produces consistent and efficient estimators of the structural parameters. It is important to note that it is not necessary for the background variables to be orthogonal nor that the total sample be randomly selected from the population. The only assumption is that all sampling units that belong to the same subpopulation are randomly drawn from that subpopulation.

The model described here is also applicable to incomplete designs because by absorbing the measurement model in (2), an implicit equating procedure is applied to the response pattern. Hierarchically ordered models with interactions, for example, can be tested with a likelihood-ratio test. In the DNAP the model was successfully applied with eight background variables which partitioned the total population into 192 subpopulations.

Computer programmes

Two computer programmes were developed for the application of the measurement and structural model. The programme OPLM (One Parameter Logistic Model) uses the CML method to estimate the item parameters for the generalized one-parameter logistic model and produces four goodness of fit tests per item. Incomplete designs with a maximum of 70 different test forms and 1000 items can be handled by the programme. Version 3.0 which becomes available by october 1990 has been extended to analyse polytomous items as well (with up to 10 (ordered) categories per item).

The experimental programme SAUL (Structural Analysis of Univariate Latent variables) estimates the parameters of univariate linear structural models in incomplete designs in which the dependent variable is not observed. The responses are multivariate binary variables which can be described by the one-parameter logistic model.

References

Glas, C.A.W., 1989, *Contributions to Estimating and Testing Rasch Models*, Nijmegen, Benda.

Rubin, D.B., 1976, *Inference and Missing Data*, in: Biometrika, 63, pp. 581-592.

Verhelst, N.D., & Glas, C.A.W., *A Class of Model Tests for Multinomial Models Belonging to the Exponential Family*, (in preparation).

Select bibliography

Eggen, T.J.H.M., 1989, *Neue Verfahren bei der Kalibrierung von Messkalen (Innovative procedures for the calibration of measurement scales)*, in: K. Ingenkamp & W.H. Schreiber (Hrsg.), Was wissen unsere Schuler? (What do our students know?), Weinheim, Deutscher Studien Verlag.

Eggen, T.J.H.M., 1989, *Itemcalibratie bij vertakt toetsen (Item calibration in branched testing)*, in: W.J. van der Linden & L.J.Th. van der Kamp (Eds.), Meetmethoden en data-analyse (Measurement methods and data analysis), Amsterdam, Swets & Zeitlinger.

Verhelst, N.D., 1989, *Informatiewinst bij vertakt toetsen (Gain of information in branched testing*, in: W.J. van der Linden & L.J.Th. van der Kamp (Eds.), Meetmethoden en data-analyse (Measurement methods and data analysis), Amsterdam, Swets & Zeitlinger.

Verhelst, N.D., & Eggen, T.J.H.M., 1989, *Psychometrische en statistische aspecten van peilingsonderzoek* (Psychometric and statistical aspects of national assessment programs), Arnhem, Cito.

Wijnstra, J.M. (Ed.), 1989, *Balans van het rekenonderwijs in de basisschool*, (Achievement of elementary school students in Mathematics), Arnhem, Cito.

15 The Rationale of a School-Based Centralized Examination System

W.H.F.W. Wijnen

Introduction

The progress test, is – at least in my opinion – a new development in examination procedures with promising and interesting characteristics. A fair description of the progress test ought to include some information on the context which gave rise to the development of this test. The progress test was developed by the University of Limburg in Maastricht in connection with a new educational system adopted by this university. Some information on the University of Limburg and its educational system will serve as a backdrop to a description of the progress test and its features.

University of Limburg

The University of Limburg is a Dutch university of recent date. Officially, it was founded in 1976, but its history goes back to the latter half of the sixties. At that time, it was considered necessary to expand the facilities for medical education in the Netherlands. There was such a dramatic increase in the number of students of medicine that the seven existing medical schools were unable to accommodate them. In those days there were no restrictions to entering a medical school and speculation on demographic development led to the expectation that the population would continue to grow rapidly. This would, in the years to come, result in an increased demand for medical practitioners. On the basis of this expectation, the government decided to establish a new medical school. As it was thought undesirable for this school to function in isolation, it was decided that it should be made part of a university.

As soon as this decision was made, a struggle broke out among a number of cities, each of which wanted to have this new medical school within its boundaries. Three arguments were advanced by Limburg, the southernmost province of The Netherlands, in support of the candidacy of Maastricht:

- From the beginning, it was emphasized that a more even distribution of academic educational facilities over the country was necessary. As young people from the Maastricht region did not have an opportunity for academic education within a range of at least one hundred kilometres, an equal geographical distribution of university facilities was a first argument in favour of Maastricht.
- A second argument was related to coal-mining. The southern part of the province of Limburg had, for a long time, been completely dependent on coal-mining. In the mid-sixties the government decided

– for economical reasons – that the pits were to be closed. This meant unemployment for many people. The arrival of a university was seen as an important compensation.
- There was also a national-cultural argument related to the geographical location of Limburg between Belgium and Germany. People were optimistic about the rapid unification of Europe. It was considered to be of major importance that a Dutch institution of higher education should be established in Limburg, the region where the French, German and Dutch linguistic and cultural borders meet.

These three arguments turned out to carry enough weight. In February 1970, the Dutch parliament decided to establish the eighth school of medicine in Maastricht "as a part of the university to be established there". In 1970 the Minister of Education installed a committee to prepare the actual start. In 1972 this committee published The philosophical principles of the Maastricht School of Medicine. This fundamental philosophy contained the framework within which the development of the university took place.

September 1974 was a very important month for, at that date, fifty students commenced their medical studies at the University of Limburg. In the meantime a limit had been set on the number of students admitted to medical schools each year and opposition against the foundation of a new university in Maastricht had been gathering momentum. However, by swiftly starting a number of different courses, the opposition was outmanoeuvred. Since it would be unwise to disregard the interests of an ever growing number of students attending these courses, it became increasingly more difficult to sustain objections to the foundation of the university. So, in 1980 a School of Health Sciences began training students in social health sciences. Since then, this faculty has developed seven different programmes in this field. In 1982 Law School opened its doors admitting one hundred students in the first year. Finally, in 1984, a School of Economics was added. Presently the University of Limburg is made up of these four faculties. The number of students is about 6,000.

Principles of the educational system

As I have indicated before, from 1974 onwards there has been a student quota in medicine. Demographic grounds, on which the establishment of the new School of Medicine had been based, were replaced by qualitative arguments. The main argument in favour of the foundation was developed in the above mentioned 'fundamental philosophy' of the faculty: an educational programme that stepped outside traditional bounds. More or less in contrast with traditional medical education, the following four principles were adhered to by this basic philosophy:
- *Problem-based learning* – The student's learning process is, as far as possible, stimulated by carefully described concrete assignments. These assignments are derived from the field of professional experience. In working on them, either individually or in a group of fellow students, the students have to collect the necessary information

from various disciplines themselves. In this way the educational programme is not structured on the basis of the separate disciplines, but through the assignments offered to the students as a point of departure and a stimulus for their study, it is structured in terms of particular themes. For example, in a particular period lasting about six weeks, medical students concern themselves with problems that may arise in connection with the theme of 'loss of blood' or 'headache'. The most important aspect of this approach is, that, in working on assignments and collecting information, the students are steadily expanding their knowledge without having to resort to any excessive memorizing or cramming. Students work on assignments and study a number of subjects that can help them to make progress. Knowledge that is required for this purpose and acquired in this way appears easier to retain than information that is learned and stored without any clear need for it.

- *Self-directed learning* – Teaching by teachers is more or less replaced by learning by the students. The learning activities of the students themselves constitute the core of the educational programme. Books, articles, audio-visual aids, videotapes, and so on, are made available in order to stimulate the learning activities of the students. This form of education makes considerable demands on the students' ability to direct their own work. The students will have to manage and organize their own learning process, making use of the opportunities for learning offered by the faculty. Consequently, they themselves have to bear responsibility for what they are going to learn. Students cannot attend lectures that provide ready-made solutions to problems. Lecturing is rather the exception than the rule. It is not the teacher who, on the basis of lectures and a compulsory reading-list, determines what the students have to do; students decide themselves what is of paramount importance in working on a set of assignments, they set priorities themselves and make their own choices. The final objective of the programme is fixed, but the students determine what path they are going to follow in order to achieve these final goals. Because finding a particular answer will often give rise to new questions, requiring further insights, the students will have to decide what is to be done there and then and what will have to be studied later, in another connection.
- *Attitude-development* – Anyone who enters a profession after completing his studies is not only governed in his behaviour by the learning he has acquired in the course of his studies. He or she is also influenced by all sorts of ideas, feelings and predilections acquired since childhood. If an institution regards itself as responsible for the development of the student's attitudes something has to be done about it within the programme. This does not necessarily mean indoctrinating the students in order to teach them the desired attitudes, or to get rid of undesirable ones. The development of attitudes as part of the subject matter just implies that opportunities are built into the educational system which allow the students and the teachers to recognize their own attitudes and to become aware of the basis of these attitudes, so they can be discussed and controlled, in sofar as this is possible at all.

Attention can be drawn to these attitudinal aspects in several ways. The manner in which students relate to each other in small group sessions can be discussed and students can learn to observe their own behaviour, and that of others, and to give their opinion on these observations. It is important to emphasize that education cannot restrict itself to the cognitive elements of learning; attitudes have to be included.

— *Continuous assessment* – The objective of education cannot be restricted to providing students with an opportunity to perform more or less brilliantly in the annual examinations. A system of regular, continuous assessment would seem to be more in agreement with the objectives of education. For this reason, the educational programme of the medical school is divided into units of six weeks each. In each period the students concentrate on a specific theme or topic, for example 'loss of blood' or 'headache' as mentioned before. The topics are not chosen for any close bearing on a particular discipline. To most assignments there are aspects relevant to various disciplines. Originally, the evaluation of the learning results of the students was intended to be based on a series of unit tests. At the end of a learning-unit an achievement test – based on the expected learning results in that unit – was administered to the students. In the Dutch educational system it is not unusual to make a decision about the promotion of the students to the next class once a year. Six to seven different measurements in six to seven different units – nicely spread over the season – was a first attempt to accomplish what was meant by continuous assessment.

The four concepts of the University of Limburg philosophy – problem-based learning, self-directed learning, attitude development and continuous assessment – are the corner-stones of its educational model.

Description of an evaluation problem

Within two years it became clear that the evaluation method did not match the educational philosophy which had been adopted. The educational programme was meant to stimulate the learning activities of the students. However, the work load was not equally divided for multidisciplinary problems can be solved through a variety of learning activities. A test that is closely related to the content of a learning-unit is – in that case – not the most obvious measuring device. Contrary to the intentions of the educational principles, the students became increasingly test-oriented. Instead of stimulating their motivation and interest, the evaluation method pushed them to selecting learning activities that led most directly to success, and induced students to choose learning activities most intimately related to the tests that were to be administered.

The discrepancy between the educational principles and the approach to evaluating the learning results had to be reduced. At the time, the medical school was not prepared to reconsider the educational principles or to draw

up a different educational programme. Principles and programme were regarded as valuable and effective and therefore it was the evaluation problem that had to be addressed. In other words, the problem was how to develop an evaluation method which allows for the selection of different learning activities by students on the one hand, and which gives a fair estimation of the results of these different learning activities on the other.

As a possible solution a progress test was developed. Progress tests can be described as comprehensive achievement tests which are independent of specific courses within an educational programme. Instead, they are related to the envisaged objectives of the programme as a whole. In a medical school this means, that progress tests are designed to ascertain whether or not student-physicians have qualified themselves at the end of their basic training.

Detailed description of a progress test

In somewhat greater detail, progress tests can be described as follows: progress tests provide a longitudinal assessment of the progress towards the final curricular objectives in the cognitive domain. This means that, whenever a certain test is administered to students of all years, the mean result of a higher year-group, say year four, must be significantly higher than the result of the year-group below, year three (and this, in turn, higher than that of year-group two, etc.). This means that in a period of about six years students have to attain a level that qualifies them . A combination of several progress tests will indicate to students how far they have advanced towards the final goal. It has to be stressed that progress tests are restricted to the cognitive domain. Some further characteristics of progress tests are the following:

- The items in a progress test are of the true/false format with the added option of a question mark. Ease of construction and the possibility of selecting more elements within the same field led to a choice in favour of the true/false format. The question mark was added to prevent guessing. The test score is determined by the number of items answered correctly minus the number of items answered incorrectly.
- The number of items per test is in between 250 and 300. Such a number of items makes it possible for progress tests to cover a particular field of medical discipline. The parallelism between different progress tests is fairly high.
- The items are selected from an item bank containing about 14,000 items. The decision to minimize the use of the same items in different progress tests resulted in an item bank of this magnitude within a period of 13 years. A good reason for not using items more than once was the fact that progress tests are also used as learning aids (students keep the test-booklets, the key to the test as well as references to the literature where the correct answers can be found).
- The progress tests are constructed on the basis of a blueprint that is subdivided into 15 organ system categories. The cardiovascular system and the respiratory system are two examples. Each progress test is

constructed anew according to this blueprint based on the International Classification of Diseases.

- The item bank contains items from about 30 medical disciplines or disciplines that are relevant in a context of medical training. A hypothetical basket is regularly replenished with new items constructed by members of the different departments. Within the boundaries, set by the blueprint mentioned above, new progress tests can be made up out of random selections of items from this basket.
- The selection of items from the basket guarantees a sample of all areas of medical science.
- Progress tests are attuned to graduate level, that is, the items are at qualification level.

About the administration some additional information can be given:

- All students are submitted to the same test at the same time. This means that all students, irrespective of the class they are in, have to take the same test. The results expected and achieved will differ substantially from class to class and also within classes. These results are only natural as an instrument is used which aims to measure the accumulation of knowledge in the course of a curriculum as a whole.
- A progress test is administered four times a year resulting in 24 points of measurement throughout a curriculum lasting for six years.

Figure 1 Example of test results of fifteen organ system categories

Indentification number: 1

Year group:

Progress test: Mar 1984

***

— About 50 general practitioners serve as a reference group. This reference group was necessary for the development of progress tests because it was the only possibility to evaluate whether proposed criteria for passing scores were adequate.

Usefulness of the results of a progress test

When the test-taking session is finished, the key of the progress test is given to the students. On the basis of this key they can make a first estimate of the results they achieved. It is only a first estimate because changes in the progress test are still possible as a result of students' comments, unexpected statistical findings for specific items, unintended errors in the text of the items, and so on.

If the progress test has its final form, the information given to the students is much more specific. The overall result on the test – specified in percentage correct, percentage incorrect and percentage question mark – is subdivided into the 15 organ system categories specified by the blueprint.

Figure 2 Example of test results for 31 disciplines, related to parts of a progress test

Indentification number: 1

Year group:

Progress test: May 1984

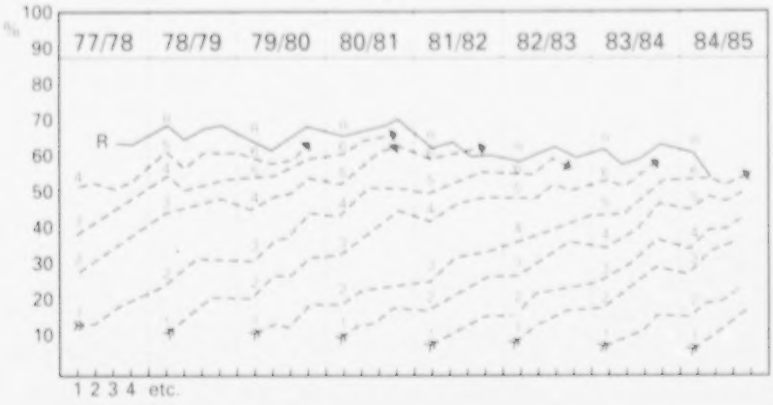
***

The same overall result and the results of the 15 organ system categories are compared with the mean results of the students who are at a similar stage in the curriculum. This comparison enables the students to discover what part of the subject-matter they sufficiently master, and on what part they should concentrate their efforts. The overall result is also subdivided into the results for about 30 disciplines, which are related to parts of the progress test.

The student can identify disciplines in which he achieved satisfactory results, and disciplines in which he did not. These results can also be compared with the means of the results achieved by students who are at a similar stage in the curriculum. They can also be used by the departments and representatives of the discipline. Thus, it can be established how different disciplines score in certain phases of the curriculum. The way in which results differ can be caused by the difficulty of the selected items, by the place of the discipline in the different units, by the opinions of the students about the importance of the discipline, and, indeed, by many other factors.

It is important to notice that these hypothetical explanations can be studied time and again, at least four times a year. They can even be studied in relation to other disciplines because all disciplines are measured on the same scale. If a series of progress tests is available, some new information can be visualized.

Figure 3 Maastricht Progress Test '77-'85, mean correct score per year-group



The graph shows the mean correct score per year-group from 1977 up to, and including, 1984 (32 administrations in all). In 1977 only 4 year-groups took part in the programme. From 1978 to 1985 a national sample of recently graduated physicians also completed each test serving as a reference group. The top shaded line represents their average scores.

As has been explained before, the reference group was included to have some estimate of the performance required of students before final graduation, and to assess the quality of Maastricht graduates. The reference

group also rated each progress test item for its relevance as to what a qualified physician should know. This has been used for item-analysis to assess test-quality. The graph shows that different year-groups achieve different scores and that the achievements of each cohort steadily improve to reach a final level nearly approaching the performance level of the reference group. Such comparisons can also be made with respect to organ system categories or with respect to the various disciplines. It is possible to trace the way in which knowledge accumulates throughout a curriculum.

Advantages of a progress test approach

The example described thus far refers to a medical school which is part of the higher education system. But the advantages of the approach are not restricted to medical schools or to higher education. That is the reason why this approach is presented at this conference. Now, what are these advantages?

- *Minimal steering effects on student learning* – Progress tests are comprehensive, they cover substantial areas of knowledge. Therefore it is impossible for a student to prepare himself in a specific way. Elements from all disciplines will be part of the test and a specific preparation is almost impossible.
- *Individual learning activities are rewarded* – The learning activities of students are not only stimulated by an educational programme. Reading journals and newspapers, watching TV-programmes and videotapes, health problems or questions about health in one's own environment can stimulate learning activities in a very effective way. The effects of these learning activities will influence the scores on the progress tests.
- *Focus on relevant, more long term knowledge* – It is of no use for students to memorize subject-matter a few hours before taking a test because those sorts of items are not included in progress tests. The items that are included are such that qualified physicians can answer correctly at any time. For teachers, it is difficult to adapt to the new requirements set down for item construction.
- *Recurrent testing of the same subject-matter* – All disciplines are tested in all progress tests. If knowledge is stored in such a way as to be easily forgotten, a wrong strategy is followed in relation to progress tests. By testing the same subject-matter time and again, the rapid loss of knowledge is prevented.
- *Retakes are no longer necessary* – In traditional assessment systems, resits of tests on specific parts of the programme are common. In fact they make educational programmes less effective. Progress tests do without retakes because every progress test in fact is a retake.
- *Compensation of scores over disciplines* – Not all teachers are convinced that compensation over disciplines is an advantage, because they think it allows students to skip specific disciplines. In actual practice, students do not skip disciplines because compensating for them by means of other disciplines is very hard.

- *Early detection of high scoring students* – In traditional testing programmes, characterized by series of successive hurdles the student has to take, it is impossible to identify students who are able to complete the programme in a shorter time. Progress tests provides the opportunity to identify students performing at a level not expected until a year later.
- *Rich source of information (sub-tests/longitudinal)* – Talking about the usefulness of the results of progress tests, some elements of this rich source of information have been mentioned before. The results available for analysis are those relating to the various disciplines and to the different organ systems, those achieved at the various year-group levels, and by the individual students at different moments. They can be studied in relation to each other.
- *Excellent research possibilities* – Relations between disciplines within a curriculum can be studied because the learning results of all disciplines are measured with the same instrument. Comparisons between year-groups or classes within the same curriculum can be studied because all students of all classes repeatedly take parallel tests. Graduates from different medical schools can be compared because progress tests are curriculum- independent. The effectiveness of different parts of the curriculum can be analyzed in a promising way. Studies of the development of cognitive abilities can be extended to studies regarding long-term effects.

This shortlist gives an impression of research possibilities based on the results of progress tests.

Disadvantages of a progress test approach

Advantages are not the only characteristics of progress tests. Some disadvantages are summed up below. This list is substantially shorter, but this conference might well add to it. So far it features the following:

- *Long testing time required* – Especially for students at the end of their curriculum, test-taking time is long. They have to answer most if not all the items and a progress test of 250 to 300 items almost four hours to complete. However, the fact that progress tests replace a series of programme dependent tests might make this long testing time more acceptable.
- *Unreliable profile scores* – The reliability of progress tests as such appears to be acceptable: based on the entire student population (all six years) Cronbach's alpha ranges from .98 to .96 for the correct score and from .91 to .85 for the correct minus incorrect score. The reliabilities within separate year groups are lower but still acceptable (Vleuten, C. van der, et al., 1990). Although reliabilities for the progress test as a whole are acceptable, the reliabilities for various sub-tests can be very low, depending on the length of the sub-tests considered. Using sub-tests as a basis for decisions about the promotion of students within a curriculum would not be justified. Sub-tests are,

therefore, only used in a formative way. Careful interpretation of these sub-scores is necessary.

- *Application to less efficient test-item formats is difficult* – All items in a progress test have the true/false format with the added option of a question mark. Other test-item formats can not be used with the same degree of efficiency. Many true/false items can cover a substantial part of the subject-matter satisfactorily whereas multiple-choice items do not get further than half the sample of true/false items within the same testing time. Some teachers don't like the true/false format. They are convinced that their discipline cannot be measured properly by true/false items. Nevertheless, for many teachers this test-format is acceptable.
- *Possibly less sensitive in freshman years* – In the Dutch system of higher education, important decisions have to be made at the end of the first year of study. It is decided whether the student will be able to complete a study at university level. Within the medical school of the University of Limburg, there has been some discussion about the usefulness of progress tests in this decisive situation. Research by Imbos (Imbos, Tj., 1989) made clear that the use of progress tests is acceptable in the first year. An exception must probably be made for low scoring students.

The progress tests are only one element in a more complicated assessment system.

Progress tests within an assessment system

In the assessment system of the medical school of the University of Limburg, there is a distinction between four different competencies: knowledge, skills, problem-solving and attitudes. These competencies are – as far as possible – measured separately. Knowledge is measured by a series of block tests which are still part of the assessment system, and by the progress tests which were described more extensively in this presentation. Skills are measured in skills tests.

A skills test is administered in all classes once a year. A single test consists of between six and twelve stations selected on the basis of a blueprint containing categories such as neurological skills, gynaecological skills, laboratory skills, interviewing skills, etc. Together, these stations form one circuit of two hours. An individual student rotates through the circuit demonstrating a different skill in each station. Performance is registered by an observer, a faculty member who is generally a specialist in the subject tested in that particular station and specifically trained for the role. The observer scores the performance on detailed checklists, on which the required task is operationally itemized. The checklists are based on standards that are used for training the students. Problem solving is measured by clinical ratings and structured oral examinations. Sometimes procedures which are specific for some disciplines are added.

The measuring of attitudes is still an unsolved problem. At this moment there are no formal assessment procedures. Students do have encounters with patients (that is, actors posing as such); they have to work in tutorials with a group of fellow-students, and are involved in so-called attitude discussion groups. It is hoped that these activities are important for the development of students' attitudes, but until now there has not been a formal assessment procedure.

Concluding remarks

The experiences with progress tests in a medical school are very promising. They offer many opportunities for studying an educational system comprehensively. The use of progress tests is not restricted to medical schools as can be shown by examples from the school of health sciences, and the school of economics. These schools too have gained experience with progress tests. Their use need not be restricted to higher education. This is a conviction which cannot be supported as yet by concrete examples. There are no reasons to expect that the basic philosophy behind progress testing should not be applicable to other parts of our educational system.

References

Barrows, H.S., Tamblyn, R.M., 1980, *Problem-Based Learning – An Approach to Medical Education*, New York.

Imbos, Tj., 1989, *Het gebruik van einddoeltoetsen bij aanvang van de studie* (The use of end objective related achievement tests in freshman years), Maastricht, University of Limburg (Doctoral Dissertation).

Schmidt, H.G., 1983, *Problem-based Learning: Rationale and Description*, in: Medical Education, 17, pp. 11-16.

Schmidt, H.G., De Volder, M.L. (Eds.), 1984, *Tutorials in Problem-Based Learning*, Assen, Van Gorcum.

Verwijnen, G.M., e.a., 1982, *The Evaluation System of the Medical School of Maastricht*, in: Assessment and Evaluation in Higher Education, 3, pp. 225-244.

Vleuten, C.P.M. van der, 1989, *Naar een rationeel systeem voor toetsing van studieprestaties in probleemgestuurd medisch onderwijs* (To a rational system for achievement testing in medical problem-based learning), Maastricht, University of Limburg (Doctoral Dissertation).

Vleuten, C. van der, Wijnen, W. (Eds.), 1990, *Problem-Based Learning: Perspectives from the Maastricht Experience*, Amsterdam, (Thesis).

16 . Measuring Educational Progress with a Longitudinal Testing System

T. L. Willoughby, A. R. Bixby,

Introduction

This chapter describes a longitudinal, comprehensive examination, Quarterly Profile Examination (QPE), used in the U.S.A. for formative evaluation since 1972. The QPE represents the second level of a three-tiered system which balances the competing needs for internal and external assessment. The three levels of assessment include: course-specific, institutional, and national. These levels of assessment address the objectives of the course, institution, and nation, respectively. Development of an institutional examination based on the curriculum of that institution provides a mechanism for assessment useful to students, teachers and administrators. In this open system, the students have access to the test item library for self-testing and receive thorough feedback from thee QPE's. Results and benefits will be described and resources required for implementation will be discussed.

The system described in this chapter is used by a medical school in the U.S.A. (Dimond, E. G., 1986) which confers both baccalaureate and M.D. degrees at the conclusion of the six-year programme. It should be noted that a similar system for longitudinal measurement has been development at the University of Limburg School of Medicine in Maastricht (Albers, W., Does, R. J. M. M., Imbos, T., Janssen, M. P. E., 1989). Furthermore, the concept has been applied to other programmes in The Netherlands. Although the QPE was implemented in a post-secondary institution, the system of longitudinal measurement can be adapted for use in a secondary school setting. The primary features of the examination system include:

- longitudinal assessment;
- comprehensive content coverage;
- student access to test item library;
- thorough feedback to students;
- computer-assisted implementation.

This chapter will focus on the educational issues and benefits of the system rather than the technical aspects of the test item bank.

Historical perspectives

Aspects of educational measurement were related to the development of the approach described herein. The influence of longitudinal testing, criterion-referenced testing, computer-assisted test construction (CATC), and value-added testing, warrant consideration before describing specific experiences at the University of Missouri – Kansas City School of Medicine (UMKC).

Computer-assisted test construction

Wood and Skurnik described the application of computer technology to the task of test generation in 1969 (Wood, R., Skurnik, L. S., 1969). By the mid-1970's, the use of computers for generating tests was widespread. These computer-assisted test construction systems generally included the following elements:

- a pool of test questions;
- an indexing scheme for retrieval;
- test generation routines;
- maintenance routines.

Test items were stored on disk in a 'Test Item Library'. Generally, user-defined classification codes were applied to individual test questions in order to provide retrieval by these codes. Some applications used existing coding schemes, such as the Medical Subject Heading (MeSH) codes developed by the National Library of Medicine in the U.S.A. Given a test blueprint, specific item numbers or classification codes, the computer would produce a test and answer key. Typically, these CATC systems were used for production of tests covering a wide range of subjects. For example, a school district or university would maintain a large pool of questions to serve the needs of numerous teachers. The faculty member could submit a search request for the generation of a quiz, give the quiz, submit the answer sheets for scoring, and have the item statistics recorded in the test item library. Existing pools of test items represent a wide range of subject matter. However, it is infrequent that the test item library is designed to cover a specific, comprehensive domain of content. The test item library at UMKC School of Medicine, computerized in the Summer of 1973, was an ambitious project at that stage of development in the CATC field. A period of three years was required to implement the full range of features.

Criterion-referenced testing

The shift from norm-referenced to criterion-referenced testing was likely the most dramatic change to occur in the field of educational evaluation and measurement during the decade of the 1970's. As Glaser (Glaser, R., 1963) pointed out, the norm-referenced approach compares the individual's score with that of other students; however, the criterion-referenced approach compares the individual's score with a previously defined standard. Popham and Husek (Popham, W. J., Husek, T. R., 1969) stimulated changes in the approach to test development activity in the early 1970's with their writings about practical ways to implement concepts related to the criterion-referenced approach. The criterion-referenced movement encompassed several related approaches such as, domain-referenced, objectives-referenced, competency-based, proficiency, and mastery tests. The terms, domain-referenced and objectives-referenced, are descriptive of the system described herein. The concepts and methods of domain-referenced testing directly influenced the development of the longitudinal assessment system at the UMKC School of Medicine.

Value-added testing

The focus on educational outcome in the U.S.A. was accompanied by discussion of concepts such as setting of standards, quality assurance and the responsibility of schools to educate its students. The book, In Pursuit of Degrees with Integrity, prepared by Northeast Missouri State University (McClain, C., 1984) described that institution's experience applying the value-added approach to undergraduate assessment. The following quotation sums up the value-added approach quite well, "Value-added refers to the positive influence an institution has on its students, to the gains students make toward desired goals. An institution begins by identifying those outcome goals. It then identifies points in time – from entry to completion – when a student's progress should be measured" (McClain, C., 1984, p. 15). The value-added approach to assessment has been applied in many settings with success. Although the importance of repeated assessment during the programme has been recognized, the measures employed generally were not both longitudinal and comprehensive in nature. Similar to other institutions which accepted the responsibility for measuring the educational impact on students, UMKC's School of Medicine made a commitment to provide its students a reliable and valid method for assessing progress.

Longitudinal testing

The Yardstick Project (Harding, H. R., 1975) hypothesized that "growth is the measurable value added by schools". This project began in Cleveland during 1966. One of their significant contributions was the Growth/Change Gauge which charted the growth of achievement of groups of students from year to year. The longitudinal approach to assessing the impact of education used in the Yardstick Project was a step in the right direction. Still, they were hampered by the constraints imposed by the use of norm-referenced tests. Thus, the measurement of growth was determined by comparing grade-equivalents or percentiles at successive years. Domain-referenced testing conducted in a longitudinal manner offers the potential for measurement of growth compared against a criterion (i.e., the domain of curricular objectives). By definition, assessing growth requires a longitudinal approach to testing. While a norm-referenced test could be used longitudinally to assess growth, the growth data would indicate growth compared with peers. It seems self-evident to the present authors that educational growth data should reflect progress toward specific objectives, not standing in ones peer group. Furthermore, the educational philosophy at the UMKC School of Medicine has taken a longitudinal perspective in both teaching and evaluation. Most of the educational research conducted at this institution has also taken a longitudinal approach. For example, Falchikov and Boud (Falchikov, N., Boud, D., 1989) reported our study of self-evaluation (Arnold, L., Willoughby, T. L., Calkins, E. V., 1985) to be unique among studies of self-evaluation in using a longitudinal methodology. As stated earlier, colleagues in The Netherlands have played a leading role in the development of longitudinal testing.

Development of the Quarterly Profile Examination

The initial development efforts involved the specification of goals and objectives of the instructional programme. The committees involved during this process were interdisciplinary. Consideration of the entire instructional sequence allows an individual teacher to place their instructional contribution into context. Additionally, faculty are better able to assess prerequisites and anticipate what follow-up or review experiences would be helpful to students. Curricular objectives were classified as pertaining to concepts, facts or skills. Having defined the objectives, test questions were generated to reflect the objectives related to facts. Questions were classified for computer storage and retrieval and entered into the test item library. Other evaluation methods were devised to assess concepts and skills. In summary, the faculty developed a comprehensive statement of the programme's objectives. These objectives specified the domain of content to which the subsequently developed test item bank should reference.

Below, the steps involved in the development of the Quarterly Profile Examination are indicated. The general principles and methods involved in linking testing to objectives were described well by Nitko (Nitko, A., 1989) in Linn's *Educational Measurement*.

1 Define the objectives of the curriculum

2 Write questions to match the objectives

3 Index and enter questions into the item bank

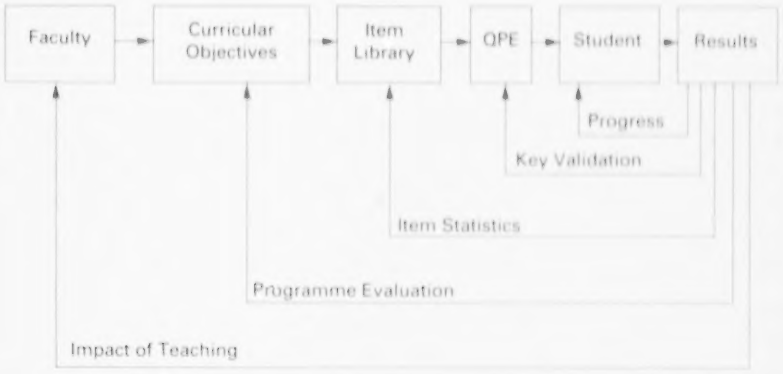
4 Develop a test blue print

5 Create an examination

Each QPE consists of 400 multiple-choice questions covering the areas of Internal Medicine, Paediatrics, Obstetrics/Gynaecology, Surgery, and Basic Sciences. The examinations are administered during the months of february, june, august, and december to students at all levels of the programme. The figure on the next page illustrates the role played by the faculty in specifying the objectives of the curriculum, which in turn determined the content of the test item bank and the Quarterly Profile Examination. Also, illustrated in the diagram are the varied functions served by the feedback derived from the QPE results.

Implementation of the test item library and testing system was initially conducted using a mainframe computer system. Users were provided access via remote terminals and printers. Advances in microcomputer technology make it possible to implement the system on a microcomputer local area network. The technological changes have made it practical to provide students with self-test disks, for use on microcomputers in addition to terminal access. Routine examination reports include individual and group test results, including an indication of items missed and references for each question. The reports include scores for 25 categories of the basic and clinical sciences of medicine.

Figure 1 *Interrelationship of faculty, curricular objectives, evaluation, and feedback*



A summary of QPE results

Analysis of QPE results indicate that the students average performance ranges from approximately 25% at year 1 to 75% at year 4 (see figure 2). The pattern of growth matches fairly well the pattern of courses taken. In particular, the results document the success we have had in teaching the basic and clinical sciences in an integrated, interdisciplinary manner.

Analysis of QPE data pertinent to specific courses has proven very useful for determining the impact of teaching. Consider the following example. During year 3, one-half of the class takes Anatomy during the Summer and the other half takes the course during the Fall. The QPE scores document the positive impact the anatomy instruction had on the students' knowledge. The cohort of students taking anatomy during the summer

Figure 2 *Longitudinal QPE performance for the class of 1990*

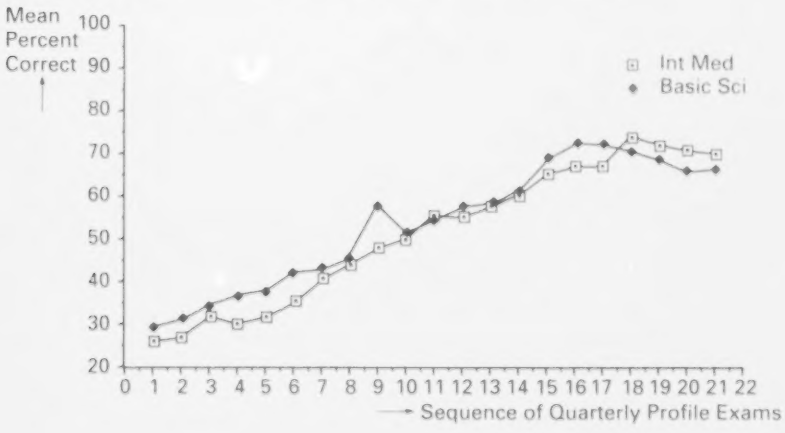
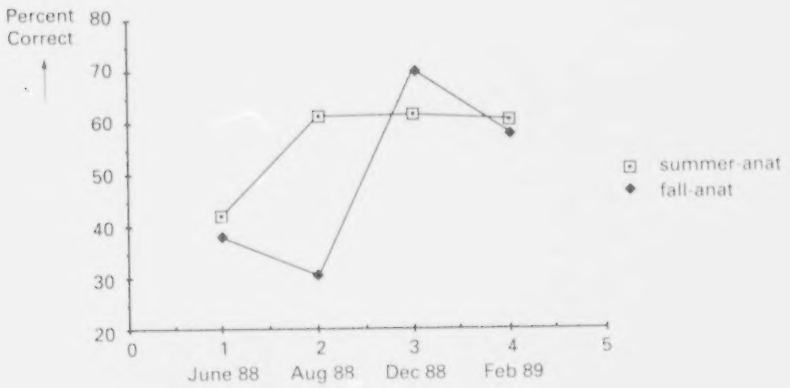


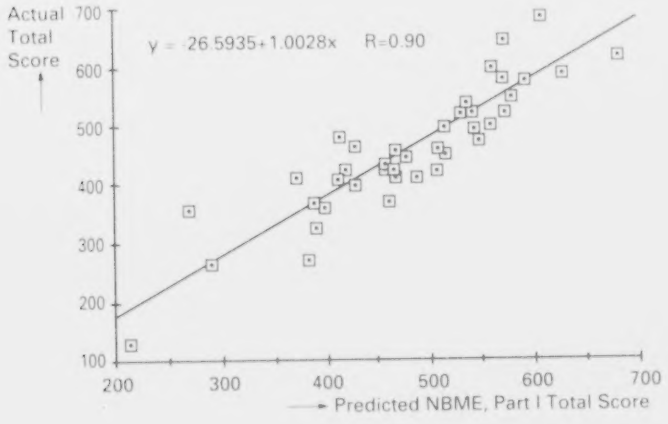
Figure 3 *Anatomy QPE scores for summer and fall rotations*



showed a substantial increase from June to August; meanwhile the other cohort of students showed a decline in scores (see figure 3).

In similar fashion, the cohort of students taking anatomy during the Fall showed an increase in scores from August to December. Previous reports have considered the edumetric validity (Willoughby, T. L., Hutcheson, S. J., 1978) and concurrent validity of the Quarterly Profile Examination for specific cohorts. Other analyses indicate that the QPE correlates highly with the national standardized test in medicine. The QPE has been reported to correlate significantly with Parts I (Willoughby, T. L., Dimond, E. G., Smull, N. W., 1977) and II of the National Board of Medical Examiners examination (Hubbard, J. P., 1978). Figure 4 illustrates the correlation of scores predicted by the QPE with the actual scores obtained on NBME Part I. Similar results (values ranging from .74 to .90 have been obtained every year since 1975.

Figure 4 *Correlation of NBME scores with QPE predictions*



QPE scores have been shown to correlate significantly with clinical performance evaluation occurring during Medical School (Willoughby, T. L., Gammon, L., Jonas, H., 1979; Willoughby, T. L., Arnold, L., Calkins, V., Neaderhiser, M., 1988). The examination results have assisted students prepare for standardized examinations and helped the institution document the results of its instructional programme.

In summary, the implementation of an institution-wide testing system affords many benefits to students, faculty and administration alike. A longitudinal, comprehensive testing system based on the objective of an educational programme can serve as a balance between needs for internal and external assessment. The model described in this chapter appears to have potential for application in other settings. In particular, application of this methodology to secondary education will be discussed.

Application of longitudinal assessment to other programmes

For purposes of illustration, consider a hypothetical example where the purpose is to create a system of longitudinal assessment for secondary school. In this example, consider that the secondary school is designed to fulfil standard requirements for High School graduation and provide a groundwork for pursuit of a career in the health field. This particular example was chosen because it illustrates both the application of the QPE concepts to secondary schools and the capacity for such a system to assess knowledge related to a specific academic programme.

Typically, the requirement for graduation from an American High School are calculated on the basis of units of credit which are equated to one academic year (two semesters) in a given course. For example, satisfactory completion of a year of Algebra results in one unit of credit toward graduation. The typical requirements are summarized in table 1.

Table 1 Units required for completion of High School

Subject	Units Required
Language Arts	4
Mathematics	4
Science	4
Social Studies	4
Foreign Language	2
Physical Education	1
Art	1
Electives	2

These courses are taken during grades 9-12. Although the requirements apply to all students, individual differences are noted in both course schedules and achievement levels. State-wide standardized achievement tests (e.g., Missouri Mastery Achievement Test) are administered to all students at grades 3, 5, 8 and 10. Iowa Basic Skills Tests are given K-8 and

Table 2 Composition of CUREVAL committees

Language Arts	
Grammar	Mathematics
Composition	Science
Literature	Social Studies
Medical	Terminology
Mathematics	
Algebra	Language Arts
Plane Geometry	Science
Trigonometry	Social Studies
Biostatistics	
Science	
Physics	Language Arts
Chemistry	Mathematics
Biology	Social Studies
Human Biology	
Social Studies	
History of World Civilization	Language Arts
Public Health and Epidemiology	Mathematics
Medical Ethics & Healthcare Issues	Social Studies
Community Healthcare Preceptorship	

Tests of Achievement Proficiency are given during grades 9-12. Students elect to take college entrance examinations, particularly in grades 11 and 12. Some course examinations are comprehensive in nature; however, what is not in place is a system of longitudinal, comprehensive assessment of individual progress.

For purposes of illustration, consider the possibility of an examination, High School Profile Examination (HSPE). In particular, the proposed examination will be designed for a special High School in which students are prepared for careers in the health field. The subjects covered would include the subjects: Language Arts, Mathematics, Science, and Social Studies. The units of instruction for these areas would be tailored to the goals of the school. Language Arts would consist of Grammar, Composition, Literature and Medical Terminology. Mathematics would consist of Algebra, Plane Geometry, Trigonometry and Biostatistics. Science instruction would include Physics, Chemistry, Biology and Human Biology. Social Studies would consist of History of World Civilization, Public Health and Epidemiology, Medical Ethics and Healthcare Issues, and Community Healthcare Preceptorship. In order to develop the HSPE, the following steps are proposed:

1 Appoint interdisciplinary committees for the four areas

2 Develop statement of curricular objectives

3 Write test questions for each objective

4 Bank the questions in computer with appropriate classification codes

5 Develop test blueprint and schedule of test administration

6 Implement testing

The work of the interdisciplinary committees will be called CUREVAL to reflect the combined focus of curriculum and evaluation. The committees should include representatives indicated in table 2. Thus, each committee consists of seven representatives, four from the primary content area and three from the other curricular areas. With proper encouragement, the process of developing the curricular objectives and testing system can foster an atmosphere of cooperation among faculty.

Also, questions from individual faculty benefit from the committee discussion and are enhanced by the feedback of others. The interdisciplinary representation fosters development of a holistic perception of the curriculum. The likelihood of overemphasis of content in any one subject area is lessened, and the integration of instruction is improved.

As questions are prepared for entry into the computerized item bank, they should be keyed to specific objectives and referenced to the textbook used. The proposed test blueprint is based on a total of 400 questions administered three times per year. The typical student would take 12 examinations during High School. The profile of content is presented in table 3.

The HSPE could be scheduled for administration at the beginning, midpoint and end of each school year, September, January, and May, respectively. Thus, the examination results would provide pre-, mid- and post-test data. Additionally, the impact of forgetting during the summer months would be easily measured. The examination should be viewed as a power test and appropriate time allowed. Perhaps the tests should be given across a two day period. Whether the test is administered via paper and pencil or computer is optional. Scored locally, the test results could be available within a week. It is recommended that students be allowed to retain the test booklets and be given a key along with score results. Students interested in studying their examinations would have a textbook reference to each question. Additionally, access to the computerized item bank by computer terminal or microcomputer for self-assessment could provide students with a valuable resource. For example, a student studying for a quiz in physics should be able to select physics questions for review. In our experience, an item bank containing more than seven thousand questions minimizes the concern that students simply memorize the answers to questions.

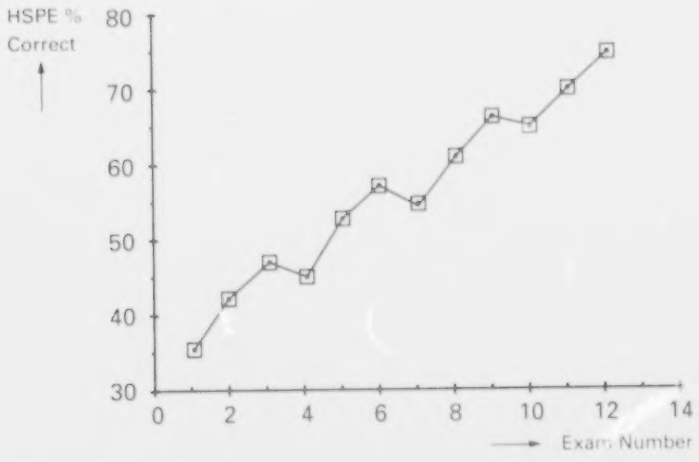
Based on the experience at the University of Missouri-Kansas City and the University of Limburg, the average score at the end of the programme is

Table 3 The test blueprint for the High School Profile Examination

Primary Subject Areas	Number of Questions
Language Arts	100
Mathematics	100
Science	100
Social Studies	100
Total	400

Secondary Subject Areas	Number of Questions
Algebra	25
Biology	25
Biostatistics	25
Chemistry	25
Community Healthcare Preceptorship	25
Composition	25
Grammar	25
History of World Civilization	25
Human Biology	25
Literature	25
Medical Ethics & Healthcare Issues	25
Medical Terminology	25
Physics	25
Plane Geometry	25
Public Health and Epidemiology	25
Trigonometry	25
Total	400

Figure 5. Hypothetical growth based on HSPE scores



predicted to be in the range of 70-80%. A growth of approximately ten percentage points per year is hypothesized. Making these assumptions, it is reasonable to project results from the HSPE. See the figure below for the hypothesized growth curve for a cohort of students advanced through High School. It is hypothesized (see figure 5) that a plateau effect will be observed for each summer period when students are not attending school.

Based on experience at UMKC (Willoughby, T. L., Gammon, L., Jonas, H., 1979) and a review of research on teacher-made tests (Hoge, R. D., Coladarci, T., 1989), the correlation of HSPE scores and teacher-made tests will be in the range of .40 to .70. It is anticipated that HSPE scores will correlate at a level of .70 to .80 with nationally standardized tests of achievement. Unlike other tests in current use, the HSPE scores provide a comprehensive assessment of growth toward mastery of the school's curricular objectives.

Benefits of using longitudinal, comprehensive assessment techniques

Faculty Benefits

With proper encouragement, the process of developing the testing system can foster an atmosphere of cooperation among faculty. Questions become public and are enhanced by the feedback of others. Also, the item analysis data resulting from use of the questions help faculty refine the item bank.

Student Benefits

Regular progress reports, including thorough feedback are provided to students in this system. Students are able to monitor their areas of strength and weakness and could be provided opportunities for self testing via remote access of the test item library. Being comprehensive in nature, the examination provides disincentive for cramming and provides an indication of content to be studied in the future.

Administration Benefits

The institution has an opportunity for collection of research data and data pertinent to programme evaluation. The longitudinal, comprehensive nature of the examination provides convincing evidence of 'value-added'. Construction of a longitudinal growth curve of knowledge is possible from this form of testing.

In conclusion, a longitudinal approach to testing has been developed and used successfully in Kansas City and Maastricht. An overview of the system and a summary of the results obtained suggests that the concepts might well be applied to academic programmes in other settings. In particular, we proposed an examination system which might be applied to the special needs of a High School programme. The concepts of longitudinal

measurement, as exemplified by the QPE, are consistent with the principles of criterion-referenced and value-added testing. The results of longitudinal testing could satisfy the public's demand for documentation of the educational impact on students. Additionally, the methods of computer-assisted test construction offer an efficient, cost-effective means of implementation. Recent reports of computer-based testing systems (Olson, J. B., Cox, A., 1990; Willis, 1990) describe systems which contain many of the elements necessary for implementation of longitudinal assessment. Our success in using longitudinal assessment and recently reported projects indicate a high probability of increased use of longitudinal testing.

References

Albers, W., Does, R. J. M. M., Imbos, T., Janssen, M. P. E., 1989, *A Stochastic Growth Model Applied to Repeated Tests of Academic Knowledge*, in: Psychometrica, 54, pp. 451-66.

Arnold, L., Willoughby, T.L., Calkins, E.V., 1985, *Self-Evaluation in Undergraduate Medical Education: A Longitudinal Perspective*, in: Journal of Medical Education, 60, pp. 21-28.

Dimond, E.G., 1986, *The Academic Plan for the School of Medicine*, Kansas City (MO), The University of Missouri.

Falchikov, N., Boud, D., 1989, *Student Self-Assessment in Higher Education: A Meta-Analysis*, in: Review of Educational Research, 59, pp. 395-430.

Glaser, R., 1963, *Instructional Technology and the Measurement of Learning Outcomes: Some Questions*, in: American Psychologist, 18, pp. 519-21.

Harding, H.R., 1975, *The Yardstick Project Interim Report*, Cleveland (Ohio), The Yardstick Project.

Hoge, R.D., Coladarci, T., 1989, *Teacher-Based Judgments of Academic Achievement: A Review of Literature*, in: Review of Educational Research, 59, pp. 297-313.

Hubbard, J.P., 1978, *Measuring Medical Education*, Philadelphia, Lea and Febiger.

McClain, C., 1984, *In Pursuit of Degrees with Integrity*, Washington, D.C., American Association of State Colleges and Universities.

Nitko, A., 1989, *Designing Tests That Are Integrated with Instruction*, in: Linn, R. L. (Ed.), Educational Measurement (3), pp. 447-74, New York, American Council on Education and Macmillan.

Olsen, J.B., Cox, A., 1990, *Development, Implementation, and Validation of a Computerized Test for Statewide Assessment*, in: Educational Measurement: Issues and Practice, 9, pp.7-10.

Popham, W.J., Husek, T.R., 1969, *Implications of Criterion-Referenced Measurement*, in: Journal of Educational Measurement, 6, pp. 1-9.

Willis, J.A., 1990, *Learning Outcome Testing Program: Standardized Classroom Testing in West Virginia through Item Banking, Test Generation, and Curricular Management Software*, in: Educational Measurement: Issues and Practice, 9, pp. 11-17.

Willoughby, T.L., Dimond, E.G., Smull, N.W., 1977, *Correlation of Intramural and Extramural Measures of Knowledge in the Clinical Sciences of Medicine*, in: Educational and Psychological Measurement, 39, pp. 897-902.

Willoughby, T.L., Gammon, L., Jonas, H., 1979, *Correlation of Clinical Performance during Medical School*, in: Journal of Medical Education, 54, pp. 453-460.

Willoughby, T.L., Hutcheson, S.J., 1978, *Edumetric validity of the Quarterly Profile Examination*, in: Educational and Psychological Measurement, 38, pp. 1057-61.

Willoughby, T.L., Arnold, L., Calkins, V., Neaderhiser, M., 1988, *The Concurrent and Predictive Validity of the Quarterly Profile Examination*, in: Educational and Psychological Measurement, 48, pp. 537-42.

Wood, R., Skurnik, L.S., 1969, *Item Banking*, London, National Foundation for Educational Research in England and Wales.

17 The Wessex Project: A Modular Development in English Post-Compulsory Education

J. Wilmut, R. G. Rainbow

Introduction

Post-compulsory education in England begins at the age of 16+, is available both in schools and in further education colleges, and has traditionally divided very sharply between a wide range of vocational courses and the academic Advanced Level examination (usually known as A Level). Vocational courses are provided or accredited by a large number of bodies, and are often closely linked to the requirements of specific occupations or industries, and many will either link to employment or incorporate substantial amounts of on-the-job experience. A Level, on the other hand, was originally conceived as the filter for access to higher education, and has generally been operated in response to the requirements of universities, polytechnics and other higher education institutions, by a number of examinations boards. Although A Level (to which an Advanced Supplementary Level – or AS Level – has now been added) is now used more widely than for access to higher education, its form has evolved only very slowly, and there is virtually no interaction between it and the vocational provision.

The term 'academic – vocational divide' has been coined to describe this situation. This divide is characterised by the existence of different accrediting and examining bodies for the vocational and academic provisions, by courses constructed in different ways, by quite different approaches to teaching and learning styles, by different approaches to assessment, and to the role of the teacher in both course construction and student assessment. The divide is sharpened because post-16 provision in schools is focused mainly on A Levels, while that in further education colleges is more concerned with vocational courses (although A Level is usually offered as well). In consequence students are obliged to make choices between an academic or vocational route when they finish compulsory education, and then amongst specific courses within each type of provision. Opportunity to mix courses is limited, and there is little chance for students to change their minds without having to start again.

This is not a coherent structure; it has grown up over many years in response to a variety of demands. National controls over what is offered are recent and partial, and courses vary widely in their structure, content and demands. The degree of specialism is generally very high, there is no guarantee that courses will provide or support common or basic skills or experiences, and students frequently lack perspectives on matters outside their immediate subjects.

Recent Educational Changes

There has been criticism of the faults of post compulsory provision for many years, with several attempts at reform. Most did not succeed, mainly because of the diversity of interests in the existing provision, all of which sought to maintain those aspects which met their needs. The changes which have taken place have generally resulted in the addition of new provisions to the old, rather than the reform of what already existed, thus exacerbating the lack of coherence.

Recently, however, there have been far more massive changes in the educational system in England, so that the pressure for reform in post-compulsory provision has become intense. Many of the changes have been as a result of government initiatives, representing a major shift from a decentralised system, in which curriculum was managed at classroom Level, and resources dispensed by Local Education Authorities (LEAs), to a far higher central government control over the compulsory phases of education, with some power over the post-compulsory sector. These initiatives, some of which have the force of law, have economic as well as purely educational aspects, and have caused teachers and others who are professionally concerned with education into a re-appraisal of what should and could be available to students. We summarise below some of these changes, with the accompanying events and pressures, concentrating on their effects upon post-compulsory education.

The General Certificate of Secondary Education (GCSE) for all students pre-16

GCSE replaced the former examinations at the age of 16+ from 1988. Its introduction was accompanied by the establishment of criteria, both general and subject-specific, within which the syllabuses and examinations must operate, with a central control over syllabus approval. A far greater degree of uniformity has thereby been introduced, although many important pieces of locally devised provision have been squeezed out in the process. It remains to be seen whether the educational provision is, as a result, richer or more valuable than hitherto, but the need to begin afresh in syllabus writing, and the requirements of the criteria, have resulted in some valuable innovation. Thus, syllabuses are generally more relevant and seek to address more directly the needs of all students, and assessment focuses more on coursework, with greater teacher involvement in many (though not all) areas. As a consequence, many students have emerged from GCSE into post-compulsory education with happier prior experiences and expectations, only to find that many courses there are not very attractive. There is frequently poor progression from pre-16 compulsory schooling to post-16 education.

Technical and Vocational Education Initiative (TVEI)

In 1983 the government introduced a major innovation into 14-18 education called the Technical and Vocational Education Initiative (TVEI). In the first

instance, this was quite separate to other changes, but it had the effect of bridging compulsory and post-compulsory education. Since 1983 TVEI has been enlarged to encompass most schools and colleges in most parts of the country. The intention was to provide, within the existing education system, new programmes which would be broad-based, include planned work experience, offer progression over the 4 years, be available irrespective of gender and ability, and encourage initiative, problem-solving abilities and personal development (Manpower Services Commission, 1984).

TVEI has evolved substantially since then. But, at a time when local initiatives were few in number, it has provided a test bed for a large number of ideas, and has contributed substantially to the broadening of the experiences of many students in the last 2 years of compulsory education. The influence on post-compulsory education was less strong and less immediate, so that, once again, high expectations which have derived from TVEI have not always been met as students have proceeded into A Level or vocational courses, although many of the latter do seek to conform more closely to TVEI criteria. It is very noticeable that some of the notions of curriculum entitlement post-16 relate very closely to what has been practised in TVEI (see, for example, Further Education Unit, 1989).

Education Reform Act and the National Curriculum

The passage of the Education Reform Act in 1988 brought a number of organisational changes to education (of which the devolution to schools of most, if not all, of their budgets is one of the most significant), and heralded the introduction of the National Curriculum, having the force of law for the compulsory years of education. This Curriculum is being developed in stages, on a subject-by-subject basis, and it will be some years before the first students emerge from it into post-compulsory education. But its shape is already having an influence; for example, the way in which the statutory requirements for Technology have been formulated will have a profound effect on the way in which this area of the curriculum is managed, incorporating substantial portions of what are, at present, Home Economics, Craft, Design, Technology, Business Studies and Information Technology. In the foreseeable future (certainly no later than 1994) students' prior experiences and their expectations will have altered yet again, as a result of the National Curriculum, and post-compulsory education will need to respond.

Assessment and recording achievement

These initiatives have been accompanied by, or have stimulated, an increased amount of activity and interest in assessment, and particularly in the exploration of models which are an alternative to, or broader than, public examinations. There is a government-funded programme of work on records of achievement which has sought to build on many earlier local initiatives, and this type of work has encouraged a process model of assessment which is incorporated into a wider view of the sorts of teaching

and learning processes which should be developed. It embodies a shift towards a student-centred view of both learning and assessment, and addresses, in the process, matters of continuity and progression within the learning experience.

Once again, there are real difficulties at the transition at age 16. Whereas the National Curriculum (into which GCSE is to be incorporated) and TVEI will probably permit these models to flourish (though perhaps with some constraints) they are not compatible with much that happens post-16, especially in A Level. The types of modular development which have utilised and acted as a trial ground for some of these ideas pre-16 were, until very recently, unknown in A Level. However, a number of vocational courses do use a modular structure though on a slightly different basis.

Changes in higher education

Social, political, economic and educational factors have forced those in higher education to review what they are doing. Funding has been short, and students for some subjects in short supply. Higher education institutions have looked again at their methods of recruitment and at their courses, and have begun to recognise that these must take greater account of the changes taking place in education up to age 18. Quite suddenly the 'top-down' influence of higher education on what happens below has become a 'bottom-up' influence on higher education, and this adds to the pressure for change in the area of transition from the compulsory sector to universities and polytechnics.

Concerns from commerce, industry, and government

There has been much concern over the last decade at the lack of trained personnel ready for employment, and at the general shortage of professional and skilled people (Confederation of British Industry, 1989). The problems experienced by higher education are a part of this wider difficulty. Criticisms have focused on several issues: a lack of basic skills in the workforce, the narrowness of specialisation in education, and the consequent inflexibility of approach which this engenders, the lack of students' understanding of industry and commerce, the number of students who fail to get any qualification either at 16 or 18, and the narrowness of the openings permitting to higher education. Much of this criticism has informed the developments of the National Curriculum and TVEI, but the lack of breadth and opportunity in post-compulsory education, the poor success rate, and the lack of essential skills and experiences remain substantial and important problems.

The issue of core skills in post-16 education and training had already been addressed by the education inspectorate (Department of Education and Science, 1989b), where there had been an acknowledgment that all courses should be structured so that they incorporate such skills, and adopt learning methods appropriate to their acquisition. Subsequently the government asked its curriculum council to make recommendations

concerning this matter, and a report has recently been published (National Curriculum Council, 1990).

In keeping with its active role in changes to compulsory education, the government has been quite closely involved in this debate. The then Secretary of State for Education encouraged substantial reform in post-compulsory provision in a speech early in 1989 (Department of Education and Science, 1989a), calling for a bridging of the academic – vocational divide, the incorporation of core skills and experiences, and courses which were more relevant to the needs of industry and commerce. The government continues to support the principles of such bridge building (Department of Education and Science, 1990).

Many of these points, with much that has been said above about TVEI, GCSE, the National Curriculum and the rest, had earlier been discussed at some length by the Higginson Committee, appointed by the government in 1987 to make recommendations about the principles that should govern A Level syllabuses and their assessment (Department of Education and Science, 1988). It was a pity so narrow a remit was given, although the Committee managed to stretch its discussions and report beyond these limited confines. The importance of the report (whose main recommendation was not accepted by the government) was in describing the desirable attributes for an A Level course of study and examination. It spoke about rigour (placing demands on students so that they can demonstrate their fullest capabilities); rewarding higher level skills, and not simply recall; motivating students by active learning and feedback on progress; breadth, balance, and depth in studies; relevance of what is learned; flexibility to allow for change and diversity. Much of this is now part of the general discourse on the nature of post-compulsory education, and it was a substantial step in opening up an area of common ground between academic and vocational provision.

The Wessex project

Throughout this period of increasingly intensive debate, the Wessex Project has sought to promote relevant and practical changes in post-compulsory provision. The Project began in 1986, with the aims of making post-16 provision more exciting, challenging and motivating, incorporating essential core skills and experiences, providing courses which are relevant to students' needs, and having breadth and balance, and sufficient flexibility to cater for many interests and for students whose interests mature and change over the period of their post-compulsory education. It has therefore sought to build bridges across the academic – vocational divide, between pre-16 and post-16 education, and between the school or college and the outside world. The Project has thus found itself wholly in step with the evolving debate over the education of 16-18 year olds, and has, indeed, been influential in informing the thinking of others. Above all, the Project has succeeded in creating courses which students are now using successfully.

The Wessex Project is a collaboration between 5 adjacent English Local Education Authorities (Avon, Dorset, Gloucestershire, Somerset and Wiltshire) and the Associated Examining Board (AEB). It was initiated by teachers and some advisory staff in these bodies, and has continued to be a grass-roots development, funded from a variety of local and national sources. Despite its local nature its influence and reputation have been national.

The roles of the participants in the Project have become less distinct as it proceeded. It has been the tradition in England that examinations boards produce syllabuses in consultation with schools, and provide appropriate examinations upon them, but are not concerned with the processes of curriculum delivery. In the Wessex Project, while AEB has certainly fulfilled these roles, it has additionally become much more involved with the whole process of course delivery and assessment, and the mechanisms for operating all aspects of Wessex courses have been discussed openly amongst all of the partners in the Project; the teachers have been fully involved in this.

When the Project began, it was clear that the area of post-compulsory education most in need of change was A Level. It will be clear from what we have said above that there was an obvious need to make the sorts of changes which later emerged in the report from the Higginson Committee (see above p.179), together with reforms which would make it more likely that the academic – vocational divide could be bridged, thus enabling schools and colleges to offer students a far more flexible provision. The Project therefore concentrated, in the first instance, on generating new A Level courses (with Advanced Supplementary, or AS, courses added later). The flexibility which was required could only be achieved by the establishment of a modular structure, of which many of the participants in the Project already had some experience.

However, the Project has not adopted a conventional modular scheme model. From the outset it has been concerned with the promotion of progression through the A and AS Level courses which it has developed, believing that learning at this level is not readily divided into discrete units which can be undertaken without due regard for what has gone before and what follows. Many modular schemes in English education have stumbled over this issue, and have only managed to resolve it by the official or unofficial imposition of a sequence on the modules. The Wessex Project has constructed its courses in two parts, having different but complementary functions. The progressive development of knowledge, skills and understanding in each A Level course is provided by its core, which accounts for 60% of the course time, and forms the stem around which the course is built; it continues for the whole of the 2 years of the course. Attached to the core, when the student is ready, are 4 modules, each worth 10% of the course time. The modules are chosen from a bank within rules attached to each core; some of the modules are very specific to a particular subject while others can operate with several different cores. All of the modules provide opportunities for in-depth studies, for student-centred work, the study of live, relevant issues and themes, and for work across the curriculum rather than being confined to a single traditional subject. It is the modules which provide much of the flexibility and choice for students.

Advanced supplementary (AS) courses are half of the duration of an A Level course, but operate at the same standard. Wessex AS courses also consist of a core, worth 60% of the course time, but therefore half of the size of a core for an A Level course. Two modules, drawn from the same bank as those for A Level courses, are attached; each is worth 20% of the AS course. Where courses of this type have been developed by the Project, they are offered as a contrast to the A Level courses; we discuss this again below.

To date the following cores have been developed, and are in use:
A Level:
- Chemistry;
- Physics;
- Biology;
- Design and Technology;
- Modern languages (French, German, Spanish);
- Art and Design;
- Geography;
- Business Studies and Economics.

The following are under development:
A Level and AS:
- History;
- Mathematics;
- English;
- Performing Arts.

AS:
- Science and Technology;
- Using French;
- Using German;
- Food Technology.

There is now a considerable module bank which has been created by teams of teachers, often working with individuals from commerce, industry and the professions. Each module is supported by extensive learning materials (student activity guides and teacher guides) which enable students to work more independently, and which assist teachers who must themselves develop new skills in order to guide and counsel their students rather than lecture. A complete list of modules would be very long; the following are a few examples of titles currently on offer:
- Working with a Foreign Language in the UK;
- Drugs, Medicines and People;
- Biotechnology;
- Environmental Health;
- Energy for Living;
- Medical Physics;
- Appropriate Technology;
- The Designer in Industry and the Community;
- The Human Figure;
- The Weather and Climate Business;

- Threatened Environments;
- A New Business Proposal;
- Multi-National Companies;
- The Firm and its Environment.

Traditionally, student attainment at A Level has been assessed mainly by external written examinations, and teachers have had minimal involvement in this process. In the current climate in education in England, it is not possible to break away entirely from this type of assessment in all subjects; the Wessex Project has therefore attempted to treat the assessment of each subject on its own merits. Where external assessment (that is, operated directly by the examination board) is required, it is part of the core. Some assessment in some cores is conducted by teachers and externally moderated. All modules are assessed by teachers and moderated by the examining board. All modules are assessed in terms of students' abilities to seek information, use information, communicate findings, although these will be interpreted in module-specific criteria.

Many modules involve students in working outside their school or college, in commerce or industry, or in relating what they are doing to the outside world. The intention is that they shall have a broader experience and be more independent in their learning. Wessex modules incorporate the sorts of core skills and experiences that have been discussed elsewhere, and which we described in the last section. In many respects the demands of the modules, while breaking new ground at A Level, are a logical extension of what students have been doing within TVEI schemes, or for GCSE pre-16.

Student results on the core and modules are arrived at separately, and reported as each component is completed. Results are awarded on the basis of criteria which are published in the syllabuses; this is a fairly unusual provision at A Level, but one which has attracted a good deal of attention, and is in step with recent developments in the implementation of the National Curriculum. The publication of results has a powerful motivating effect on students, who can find in the syllabuses the rules of combination for core and module results to see what final grade they may obtain. It is also of considerable benefit for them to be able to demonstrate actual results when they go for interviews at universities, polytechnics, in industry or commerce, and it is the experience of the Project that they are far better able to discuss their work with those interviewing them.

We have already mentioned the need for teachers to learn how to cope with the new demands of Wessex courses. To assist in this process, local support networks have been formed, and the examining board moderators work with these networks in programmes of teacher support and student moderation. The flexibility of use of modules demands a flexible mechanism on the part of the Board for dealing with this aspect of Assessment.

It is in the assessment of the modules and some elements of cores, such as groupwork, that teachers need a good deal of help, and the Project needs to undertake further development in these areas. The widest range of assessment techniques has been used, and applied in a wide variety of situations, often involving adults other than the teachers in aspects of the scrutiny of what students have done. The products of the modules range very widely, from portfolios of art work or designs to reports of

experimental work and physical artifacts. Quite often the work is presented in what are, for A Level, unconventional media.

Evaluating outcomes

The Wessex Project is a pilot scheme, confined to the schools and colleges in the 5 LEAs concerned. It is being evaluated by AEB, reporting on the certification of students in 1989, 1990 and 1991, and by staff at the University of Bath, studying the effects of the Project on teachers and students.

At this time we can offer only the most tentative conclusions from the Project. There is no doubt that the approach which has been used has fired the imaginations of many teachers and students, and generated a huge interest both within and beyond the 5 LEAs, so that the pilot is substantially oversubscribed at present. The student and teacher guides which have been produced have found a good deal of use outside the Project, where there is a shortage of good quality materials.

Students appear to react well to the demands of the courses, although this is often true of pilot schemes, and we must wait for a year or two to see how the Project settles down. The first group to receive final A Level grades in 1989 performed very well indeed, and there have been very favourable reactions to the standard of student work from a wide range of people. Much of the favourable reaction is concerned with the motivation of students that arises from their opportunities to be more responsible for their own learning, the immediate feedback they get on their achievements in modules and from the development of essential skills. At the same time there have been problems with work overloads for some students, and not all teachers have been equally successful in discarding traditional approaches to A Level. In particular, while some teachers have welcomed the opportunity to work closely with colleagues in other departments, so that the cross-curricular potential of the Project can be exploited to the full, others have resisted the pressure to move outside the confines of traditional approaches to single subjects.

The examinations pose some interesting problems for the Board. At the administrative level it has been necessary to devise procedures which accumulate results over the two year span of the course, and then operate rules of combination for the production of final grades. More significantly, participation in the Project has required the Board to become far more immersed in the operation on the scheme in schools than would generally be the case, with a correspondingly higher level of contact with teachers, and with its procedures more open to scrutiny. For most Board staff, this Project is the most radical piece of innovation in which they have been asked to be involved, and some tensions have been created as a result.

Furthermore, one of the implications of the way in which the Project has designed its assessments is that teachers will need to have a substantial role as assessors, and that these assessments are often embedded in activities which cannot readily be seen by moderators. Teachers will need to be expert assessors (a role which is demanded of them by the National

Curriculum) and moderation will become more and more of a monitoring process. Indeed, the moderators for the Wessex Project are already participating in support activities for teachers, through local networks, which should achieve this end. There is then a significant shift in emphasis away from the Board as an external arbiter to the Board as a partner in a process of assessment; this is surely a healthier role.

In addressing the issue of bridging the academic – vocational divide, all of the partners in the Project are now being challenged even further. Contacts have been made with the providers of vocational courses, and the bodies which accredit and certificate these. Substantial parts of Wessex courses have common ground with vocational courses, and we are now embarking on small scale piloting of the delivery of common courses to students who will be able to decide at a later stage whether they will finally opt for a vocational or academic certificate. This involves some credit transfer into and out of Wessex courses, and therefore the mutual recognition by two examining bodies of each other's certification. If this small-scale bridging can be operated successfully it opens up the possibility of devising structures and courses which are more truly common, although still allowing for the distinctiveness of each type of provision to be retained.

Conclusion

We have discussed aspects of the Wessex Project as a response to a set of rapidly changing circumstances in English education. It is not a static Project: it is no longer realistic to talk in those terms in Britain. One of the practical problems faced by the Project lies in persuading those in authority that organised flexibility of response within a project, accompanied by proper evaluation, is a profitable way of advancing education. Unfortunately one of the consequences of the increased centralisation of education in Britain is a reduction of flexibility and an increased risk of a concentration on methods and solutions which have been generated at the most senior levels. Wessex is also heartening because it is teachers who are generating the ideas and the response to a changing situation. This is becoming a rarer phenomenon in English education, and it is to be hoped that the Project can continue to be influential in the evolution of new patterns for post-compulsory education.

References

Confederation of British Industry, 1989, *Towards a Skills Revolution.*

Department of Education and Science, 1988, *Advancing A Levels*, London, HMSO.

Department of Education and Science, 1989a, *Further Education – A New Strategy*, Speech by Right Honourable Kenneth Baker MP, Secretary of State for Education and Science.

Department of Education and Science, 1989b, *Post 16 Education and Training: Core Skills*.

Department of Education and Science, 1990, *Speech Made to GCE Boards*, Cambridge, April 1990 by Right Honourable Angela Rumbold MP, Minister of State for Education and Science.

Further Education Unit, 1989, *Towards a Framework for Curriculum Entitlement*, London, FEU.

Manpower Services Commission, 1984, *TVEI Operating Manual*, Sheffield, MSC.

National Curriculum Council, 1990, *Core Skills Post-16*.

18 Experiments on General High School Completion Tests in China

Xuewei Yang

Introduction

In China today, the number of youngsters reaching the same age is about 24 million. In 1989, for example, there were 18.75 million students finishing 6-year primary school, 11.34 million students graduating from the 9-year compulsive education (equivalent to middle school), and 2.43 million completing the 12-year general high school. Higher educational institutions in this year were planning to admit 640,000 students.[1] This means that the percentage of youngsters being admitted to higher schools from a same aged group is less than 3 percent. Among the 640,000 college freshmen, about 30 percent were graduates from previous years high school classes. Therefore, only less than 20 percent of the high school teenagers were admitted to higher educational institutions in the same year in which they graduated. This indicates, that for high school students, to advance to higher schools is severely competitive.

In this very competitive process of applying for higher schools, several problems have occurred in high school education: morality and physical training education in high school have deteriorated; efforts have being paid to only minority of 'prosperous' students and thus cares for the majority of students have been neglected; standard teaching curricula have not been seriously carried out while high schools are concentrating and only offering subject courses[2] that students needed for the college entrance examination, resulting in premature diversification of subject courses (students in senior high schools are advised to choose either liberal art courses or science courses). Thus students' knowledge structures tend to be unbalanced and fragmented. High school teachers, in an effort to prepare their students to cope with the hardest problems likely to appear in the college entrance examinations, are introducing advanced courses that may not be acceptable for all students.

These problems have caused wide-spread frustrations and lack of study interests in those students who consider themselves 'losers' or those who did not perform so well academically. For those fortunate ones being admitted to colleges or universities, their narrow academic backgrounds

1. There are also 130,00 students in agricultural professional middle schools, equivalent to nine grade normal education. Agricultural professional high schools, vocational schools, and vocational teachers' schools, equivalent to twelve year normal education, have enrolled 1.9 million students. Finally, adult higher education have admitted 540,000 part time students.

2. The entrance examination to higher education has two categories, liberal arts and sciences. The former one includes political science, Chinese, mathematics, history, geography, and foreign languages; the latter one includes political science, Chinese, mathematics, physics, chemistry, biology, and foreign languages.

have become obstacles in their future academic advancements.

The above problems are caused by complicated factors from many sectors of our country. One reason, the tough competitions to advance to higher education that drive the high schools to go astray, is limited by our nation's present economic and social conditions and is unlikely to be resolved in the short run. The more direct reason, is the incompatibility of examination systems in high schools. In fact, China now still lacks standard examinations to assess the level of high school graduates. Consequently, the entrance examinations to higher education have virtually become a final measure of high school teaching, and the passing rate in the entrance examination has become a indicator of high school performance. As a result, the high school teaching curriculum has become so effortless that it has been replaced, in practice, by the syllabus of the entrance examination to higher education. Being a selective test, the entrance examination to higher education is designed to select 20 percent of the student population, whereas it has been used in fact as a test to evaluate the level of 95 percent of the high school graduates. Obviously, this problem is a systematic one in the educational structure itself, and by no means related to any economic or social conditions at all, and therefore, can be resolved by reforms.

The General High School Completion Test (GHSCT) programmes

In 1983, the Ministry of Education of China (now the State Education Commission) initiated a discussion on this high school test reform. Later, a committee involving 6 experts was set up to start a research program. In 1985, a report was proposed to the State Education Commission appealing for the establishment of the General High School Completion Test (GHSCT). In the same year, a pilot programme started in the city of Shanghai. Then in 1985, more programmes were introduced in Zhejiang, Hainan, Hunan, and Yunan provinces.

The above GHSCT programmes are designed to pursue the following guide lines:

- The GHSCT is a state approved and nationally recognized test, administered by provincial testing organizations.
- The GHSCT is a proficiency test designed for high school graduates, and is regarded as a major measure to assess high school education. High school graduate certificates are issued to those who passed the GHSCT and judged morally and physically eligible. The high school graduate certificates and the GHSCT certificates are to be the necessary documents to advance to higher education[3] and to seek employment.
- The GHSCT includes examinations and assessments. There are nine compulsory subjects to be examined: political science, Chinese, mathematics, history, geography, chemistry, biology, and foreign languages. Assessment courses include labour skills and laboratory

[3] Applications to higher education include applying for the general higher educational institutions, adult higher education, and taking the self-taught higher aducation certificate tests.

experiment skills in physics, chemistry, and biology.

- High school teaching curricula will be adjusted in such a manner that once a subject course is completed, the GHSCT test of the subject would be immediately given. In the first year of high school, ten subjects would be tested, second year five, and the third year three subjects.
- The high school teaching curricula would be modified. Based on this renewed curricula and contents of the compulsory subject courses, the GHSCT syllabus would be constructed. The level of the GHSCT, compared with that of the entrance examination to higher education, is relatively lower, since the GHSCT is aimed at granting completion for 95 percent of the high school students. The syllabus for the entrance examination to higher education, on the other hand, would be based on both compulsory and selective courses offered in high schools.
- The GHSCT test papers will be developed by provincial testing organizations, who will also be grading the papers as well. For provinces with a large population, objective test papers will be scored by regional testing organizations under the supervision of provincial testing institutions. The GHSCT scores are reported in a five-grade system. The selective courses are to be tested by questions designed by the region or the school and supervised by the provincial testing institution. Selective course grades are reported as either 'pass' or 'failure'.
- As the GHSCT covers subjects taught in a particular school term, the final examination of that term will not include subjects already tested in the GHSCT.
- Candidates who fail the GHSCT are provided with chances of make-up tests.
- Based on the implementation of the GHSCT, the entrance examination to higher education will include fewer subjects (about 3 to 4) only relevant to the candidates' majors and requirements of the higher institutions that have been applied for.

Results

The GHSCT programmes in the city of Shanghai and Zhejiang province have been operational for five and two years respectively. They turned out to be effective in the following aspects:

- As a major assessment measure of high school teaching, the GHSCT has enabled high schools to consider achieving high pass rates not only in college entrance examinations, but most importantly, in the GHSCT as well. They would then allocate their efforts and resources on the majority of the students rather than the few talented students. In the mean time, the GHSCT, whose syllabus meets the general level of the high school students, has inspired an overwhelming number of students, especially those who performed not so well, to improve. On the students side, they have shown more interests in their study. Consequently, there are indications that high schools where the GHSCT

was implemented are functioning in a more orderly way.

- Because of the GHSCT's coverage of all high school subjects and its pattern of giving the tests (a test is given immediately after the completion of a subject), high schools are obliged to abandon their traditional practices of dividing students into liberal arts and sciences classes. They now rather offer a full range of courses to all students. As a result, the overall teaching quality of high schools has been greatly improved and the academic standards of high school graduates have been securely maintained.
- Based on the nine-subject GHSCT, the entrance examination to higher education would include fewer subjects which are only related to applicants' majors. This new system has provided higher educational institutions with students who are better prepared in their academic backgrounds and are more likely to develop to their maximum capacity in their majors. In general, the GHSCT would provide an academically stronger supply of students for higher education.
- The GHSCT has also constructed a stronger structure for educational authorities to enhance administration.

Experiences from the GHSCT experimental programmes have indicated that education and tests are supplement each other and are two aspects of one unity. As powerful assessment measures, tests have to be designed and coordinated with the goals of education. Otherwise, tests may turn out to be destructive.

As the try-outs of the GHSCT in several provinces have shown very promising results, the State Education Commission has decided that in three years or longer times, the GHSCT would be gradually introduced throughout the country, and eventually, reforms would take place in the entrance examination to higher education and the higher education admission process.

The GHSCT has come into experimental trial for only a short period of time, yet it still has some problems in policy making and technical specifications. In addition, efforts would be needed to resolve outside difficulties such as helping the general public to get adjusted to the new test systems.

In the GHSCT experimental programmes, we have received assistance from many internationally well known testing organizations. It is our hope that colleagues from the international testing scene would be more than welcome to take part in China's test reform programmes to provide a better education and testing service for the world's largest population.

List of Contributors

Alice R. Bixby
Address: University of Missouri, 2411 Holmesstreet, Kansas City, MO 64108, United States of America

Paul J. Black
Address: Centre for Educational Studies, King's College London, 552 King's Road, London SW 10 OUA, United Kingdom

Sorel Cahan
Address: The NCJW Research Institute for Innovation in Education, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel

John S. Cook
Address: NSW Department of School Education, 35-51 Mitchell Street, North Sidney 2060, Australia

Robert K. Crocker
Address: Memorial University of Newfoundland, St. John's, Newfoundland, Canada A1B 3XB

Theo Eggen
Address: Cito, The National Institute for Educational Measurement, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

Richard Ferguson
Address: ACT American College Testing, P.O. Box 168, Iowa City, Iowa 52243, United States of America

Cameron Fincher
Address: Institute of Higher Education, University of Georgia, Candler Hall, Athens, GA 30606, United States of America

Mary Hayden
Address: University of Bath, School of Education, Claverton Down, Bath BA2 7AY, United Kingdom

John Izard
Address: ACER The Australian Council for Educational Research, P.O. Box 210, Hawthorn, Victoria 3122, Australia

Sandra Johnson
Address: Franklin College Switzerland, Via Ponte Tresa 29, 6924 Sorengo, Switzerland

Ian J. Kerr
Address: University of Bath, School of Education, Claverton Down, Bath BA2 7AY, United Kingdom

Henk P.J. Kreeft
Address: Cito, The National Institute for Educational Measurement, P.O. Box 1034, Arnhem, The Netherlands

Anton J.M. Luijten
Address: Cito, The National Institute for Educational Measurement, P.O. Box 1034, Arnhem, The Netherlands

David L. Ongom
Address: Uganda National Examinations Board, P.O. Box 7066, Kampala, Uganda

Robert G. Rainbow
Address: The Wessex Project, Holyrood School, Chard, Somerset TA20 1JL, United Kingdom

Colin Robinson
Address: SEAC, School Examinations and Assessment Council, Newcombe House, 45 Notting Hill Gate, London W11 3 JB, United Kingdom

Jeffrey Thompson
Address: University of Bath, School of Education, Claverton Down, Bath BA2 7AY, United Kingdom

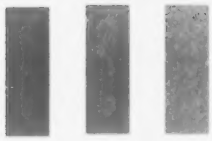
Norman D. Verhelst
Address: Cito, The National Institute for Educational Measurement, P.O. Box 1034, Arnhem, The Netherlands

John Wilmut
Address: MGB Assement Agency, 20 Brooksby Road, Tilehurst, Reading RG3 6LX, United Kingdom

Thomas L. Willoughby
Address: University of Missouri, 2411 Holmesstreet, Kansas City, MO 64108, United States of America

Wijnand H.F.W. Wijnen
Address: University of Limburg, P.O. Box 616, 6200 MD Maastricht, The Netherlands

Xuewei Yang
Address: National Education Examinations Authority, 30# Yu Quan Road, Beijing 100039, China (PRC)



UITGEVERIJ

LEMMA